Downloaded from orbit.dtu.dk on: May 06, 2023

DTU Library

Targeting the genetic complexity within adapting RNA virus populations

Fahnøe, Ulrik

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Fahnøe, U. (2014). *Targeting the genetic complexity within adapting RNA virus populations*. Technical University of Denmark.

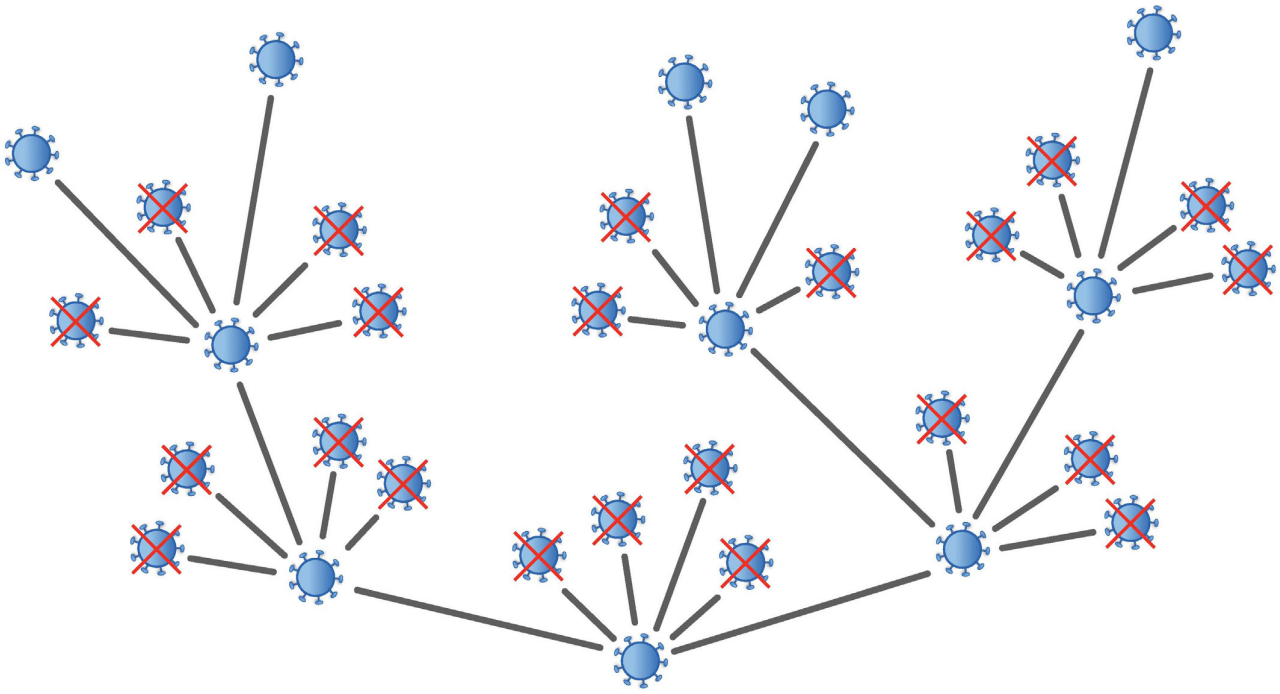
General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Targeting the genetic complexity within adapting RNA virus populations

Ph. D. Thesis December 2014 Ulrik Fahnøe



DTU Vet
National Veterinary Institute

Targeting the genetic complexity within adapting RNA virus populations

Ph.D. thesis by Ulrik Fahnøe 2014

Technical University of Denmark

National Veterinary Institute, Lindholm

Supervisor:

Thomas Bruun Rasmussen

National Veterinary Institute, Technical University of Denmark, Lindholm, Denmark

Co-supervisor:

Anders Gorm Pedersen

Center for Biological Sequence Analysis, DTU Systems Biology, Technical University of Denmark, Denmark

Cover art:

Mutating and replicating viral population, by Anders Gorm Pedersen

Table of Contents

Preface **7**

Summary **9**

Resume (dansk) **11**

Abbreviations **13**

Introduction **15**

Classical swine fever virus 15

Viral lifecycle 17

Virus population diversity 17

Molecular evolution 19

Reverse genetics 22

Next generation sequencing 23

RNA virus applications of NGS 23

NGS methods and results **25**

Sample preparation for NGS **25**

Full-length RT-PCR 26

Half-length RT-PCR 27

RNA sequencing 28

Second-strand cDNA sequencing 29

Library preparation **30**

NGS data analysis and pipeline **31**

Raw data: Fastq or SFF file 33

Read Quality assessment and trimming and filtering 33

NGS error correction and SNP call benchmarking 34

De novo assembly 36
Mapping reads to a reference sequence 37
Extraction of consensus sequence and coverage depth 39
SNP calling and SNP translation 40

Manuscripts 45

Manuscript 1 47

Complete genome sequence of border disease virus genotype 3 strain Gifhorn 47

Manuscript 2 51

Complete genome sequence of classical swine fever virus genotype 2.2 strain Bergen 51

Manuscript 3 55

Rescue of the highly virulent classical swine fever virus strain "Koslov" from cloned cDNA and first insights into genome variations relevant for virulence 55

Manuscript 4 67

Analyzing the fitness of a viral population: Only a minority of circulating virus haplotypes are viable 67

Manuscript 5 107

Classical swine fever virus adaptive response to vaccination: early signs of haplotype tropism 107

Conclusions and future perspectives 143

Other work 148

References 151

Preface

This thesis is the result of three years of scientific work and study. Most of the experimental work has been done at the National Veterinary Institute at the island of Lindholm, Technical University of Denmark. This work has also been done in collaboration with the Friedrich Loeffler institute, Germany, who performed some of the sequencing and a pig infection experiment; Center for biological sequence analysis (CBS) at DTU Systems Biology which collaborated with the molecular evolution and bioinformatics; University of Glasgow contributed further with bioinformatics.

This work has been a collaboration from the start and I could not have achieved the results without assistance from a range of people and for that I would like to share my gratitude.

First of all, I would like to thank all the people at Lindholm for their hospitality and making me feel welcome at this strange place. The three-year period I spent on the island have definitely been a great experience and I will miss the special atmosphere and the people.

I would like to thank my supervisor Thomas Bruun Rasmussen for taking a chance by hiring a molecular geneticist with no prior experience in virology or bioinformatics. I think we have achieved a lot together in the last three years and our collaboration has been inspiring and fruitful. During my PhD, we have travelled frequently together for conferences and meetings with international collaborators. Our trips have been highly productive and also fun, like our trips to Venice or Brighton to mention a few. From our group at Lindholm, I would like to highlight the assistance performed by Lone Nielsen our former technician who is now retired. Thank you for all your help and your kindness! Together we made result for the "stockpile" that is still waiting to be analysed. I would also like to thank former PhD student Peter Christian Risager for being my travel companion to and from Copenhagen and for letting me meddle in his project. A direct spinoff from Peters project was the master thesis of Jonas Kjær that I co-supervised. Thank you Jonas, for doing a great job in such short time and good luck with your PhD. My colleague Johanne Hadsbjerg is thanked for helping me with organizing Young Epizone in Copenhagen, which was a success and good luck with your PhD. Finally, I would like to thank the caretaker MarioBaltzer Petersen for taking care of the pigs; the

technicians in the culture lab, Jens Nielsen and Louise Lohse for their assistance in the pig experiments.

As mentioned above, I was and in some ways still am, a rookie in bioinformatics but luckily my co-supervisor Anders Gorm Pedersen from CBS DTU System has been there to help me. I want to thank you for your optimism and inspiration. We have done some interesting things together, especially the molecular evolution you taught me. Hopefully, we will be able to collaborate further in the future! Also thanks to Simon Rasmussen at CBS DTU System for his help with the data analysis.

During my PhD I have visited FLI in Germany three times. Their expertise has been very important for my project. For that I would like to thank Dirk Höper, Sandra Blome, Carolin Dräger and Martin Beer for their assistance and good ideas.

During my PhD I had the opportunity to go to the University of Glasgow to stay for a month in Daniel Hayden's group at the Institute of Biodiversity Animal Health and Comparative Medicine. There I worked together with Richard Orton on the sequence analysis and other bioinformatics issues. I would like to thank Richard for his excellent assistance and for taking me fishing at Loch Lomond where I caught my first Perch. I really enjoyed my stay and got the opportunity to delve deeper into bioinformatics.

At the very end I would like thank my partner Katrine for the support and your willingness to share me with the island;-)

Ulrik Fahnøe, december 2014

Summary

RNA viruses such as the classical swine fever virus (CSFV) have some of the highest mutation rates known in nature. This makes them highly adaptable to different hosts and can allow escape from host responses such as immune pressure. In addition, the high mutation rate will result in coexistence of a population of closely related haplotypes. The variation within the viral population enables the virus to adapt to changes in selection pressure and thereby evade treatment and possibly vaccination. Earlier technologies did not allow the study of virus populations in depth. However, the arrival of new sequencing technologies also known as next generation sequencing (NGS) drastically changed the sequencing capacity and thereby allowed deep sequencing of viral populations. In this thesis I deal, in particular, with the population structure of CSFV and how the virus adapts to host pressure and in what way the population affects virulence and infectivity. A fully functional and highly virulent cDNA clone of the CSFV strain "Koslov" was reconstructed and animal infection experiments revealed differences in virulence between two closely related cDNAs. In addition, deep sequencing of samples taken from the infected animals allowed us to study the molecular evolution in detail and identify adaptations in the virus populations. We then proceeded to molecular cloning of a single virus population of another CSFV strain to analyse haplotype structure and infectivity. Only a minority of the cDNA clones was functional and the infectivity was associated with the number of missense mutations. However, by reconstructing the ancestral sequences inferred by the cDNA clone phylogeny, we were able to make virus that resembled different haplotypes in the parental virus population. Infection of pigs revealed a difference in virulence between these reconstructed ancestors. I also studied virus population adaption with or without vaccination. We found that virus sequenced from immunised pigs showed more signs of adaptation than controls. In addition, we observed that a particular haplotype were more present in some part of the pig compared to another haplotype pointing to virus tropism effect.

The thesis is comprised of three parts: **Part 1**, which is an introduction to CSFV, viral lifecycle, virus population diversity, molecular evolution, reverse genetics, next generation sequencing and NGS viral applications. In addition, I have included two chapters describing NGS sample preparation and data analysis in detail. **Part 2**, the manuscripts published and/or submitted

(except manuscript 5) during this PhD. These mainly involve CSFV population studies and reconstruction of virulent cDNA clones. **Manuscript 1** describes the complete genome of Border disease virus (BDV) strain “Gifhorn” derived from a pig isolate and sequenced by NGS. **Manuscript 2** is the full-length NGS sequencing of the CSFV strain “Bergen”, which is the first genotype 2.2 strain to be fully sequenced. **Manuscript 3** describes the rescue of the highly virulent CSFV strain "Koslov" from cloned cDNA and first insights into genome variations relevant for virulence. This manuscript also includes a part that investigates the viral population molecular evolution by NGS and several interesting adaptations were observed. **Manuscript 4** describes an extensive study of a single virus population by full-length cDNA cloning and NGS sequencing. Only a minority of the cDNAs were functional and the infectivity was associated with the number of missense mutations on each cDNA clone. In addition, NGS haplotype reconstruction tools were compared to the phylogeny inferred from the individual cDNA clones. Finally, the two major haplotypes in the virus population were produced by a combination of ancestral reconstruction and reverse genetics and tested in cell culture and in their natural host revealing different phenotypes *in vitro* and *in vivo*. **Manuscript 5** describes a study using NGS sequencing of adapting challenge virus populations within immunised pigs. SNP (single nucleotide polymorphism) analysis revealed key differences between the virus populations in immunised pigs compared to the pigs from the control group. In addition, a tropism effect was observed in the control animals between the different types of samples sequenced. dN/dS analysis revealed the serum from the immunised animals to be under slightly more positive selection compared to the control animals. **Part 3**, is a chapter that sums up the findings of this thesis and discusses the future perspectives of this research field.

Resume (dansk)

RNA-virus har nogle af de højeste mutationsrater beskrevet i naturen. De høje mutationsrater medfører konstante ændringer i det genetiske arvemateriale, hvilket betyder at selve viruspopulationen forekommer som tæt beslægtede, men ikke identiske viruspartikler. Denne genetiske variation gør det muligt for RNA-virus hurtigt at tilpasse sig til ændringer i selektionspres fra for eksempel værtens immunforsvar eller fra antiviral behandling. I dette projekt har jeg undersøgt den genetiske variation i klassisk svinepestvirus og hvordan denne RNA-virus kan tilpasse sig sin værts selektionspres og i hvilken retning den genetiske variation påvirker virulens og smitsomhed. Den eksplosive udvikling i nye sekvenseringsteknologier har gjort det muligt for os at dybdesekventere en lang række viruspopulationer og derved undersøge den genetiske variation ganske nøje. Samtidig har jeg brugt adskillige andre teknikker som blandt andet involverer kloning af fuld-længde klassisk svinepestvirus, revers genetik, infektionsforsøg i grise og bioinformatik. Undervejs har jeg rekonstrueret en fuld funktionel klon af den højvirulente klassisk svinepestvirus "Koslov". Med udgangspunkt i denne klon har vi været i stand til at studere virusudvikling i inficerede grise og se på tilpasninger og forskelle i virulens mellem to tæt beslægtede viruspopulationer. Disse observationer gav os blod på tanden i forhold til at studere den genetiske variation gennem intensiv kloning af virus fra en enkelt viruspopulation. Det viste sig at flertallet af disse klonede viruskopier ikke var funktionelle i cellekultur. Yderligere fandt vi en sammenhæng mellem antallet af mutationer og hvor smitsomme de enkelte klonede viruskopier var. Ved at rekonstruere sekvensen for viruskopiernes forfædre var det muligt at fremstille fuldt funktionelt virus. Infektionsforsøg i grise viste forskelle i virulens mellem disse rekonstruerede forfædre. Jeg studerede også virustilpasning i grise med og uden vaccination. Vi fandt ud af at virus fra de vaccinerede grise viste mere tegn på tilpasning end virus fra kontrolgrisene. Samtidig så vi også tegn på at bestemte subpopulationer var mere tilstede i nogle typer væv end i andre.

Projektet har bidraget med vigtig ny viden om hvordan RNA-virus konstant ændrer sig. Viden som er meget vigtig for forståelsen af disse farlige virussers udvikling. Desuden belyser vi

også hvordan virus ændrer sig under påvirkning af vaccination, der kan bruges til at teste hvor godt vacciner beskytter mod virustilpasninger for dermed i sidste ende at kunne lave bedre og mere sikre vacciner.

Abbreviations

ADV	Aleutian disease virus
BAC	Bacterial artificial chromosome
BAM	Binary alignment map
BDV	Border disease virus
bp	basepair
BVDV	Bovine viral diarrhea virus
CSF	Classical swine fever
CSFV	Classical swine fever virus
cDNA	complementary DNA
DMAC	DTU multi-core assay center
FASTA	Fast-All
FASTQ	Fast-Quality
FMDV	Foot and mouth disease virus
HTML	Hyper Text Markup Language
Indel	Insertion deletion
IRES	Internal ribosome entry site
kb	kilobase
NGS	Next generation sequencing
NTR	Non-translated region
PID	Post infection day
PRRSV	Porcine reproductive respiratory syndrome virus
qPCR	quantitative Polymerase chain reaction
PK-15	Porcine Kidney cells
RdRp	RNA dependent RNA polymerase
RNA	Ribonucleic acid
RT	Reverse transcriptase
SAM	Sequence alignment format
SFF	Standard flowgram format
SFT-R	Sheep Fetal Thymus – Riems cells
SK6	Swine Kidney cells
SNP	Single nucleotide polymorphism
UTR	Untranslated region
VCF	Variant call format

Introduction

Classical swine fever virus

Classsical swine fever virus (CSFV) (Fig. 1) is a member of the genus pestivirus in the family *Flaviviridae*, which also contains important human pathogens such as West Nile virus, yellow fever virus and hepatitis C virus.

Infections by CSFV are restricted to pig species and causes classical swine fever (CSF), whereas other pestiviruses, such as bovine viral diarrhoea virus (BVDV) or border disease virus (BDV), infect a wider range of host animals. CSFV, in particular, has high socio- economic importance and remains a significant threat to animal health (Postel et al. 2013). A large number of CSFV strains exist which vary considerably in their virulence, including high, moderate, low and avirulent variants (Floegel-Niesmann et al. 2009). Avirulent and low virulence strains result in no or a mild disease and induce a protective immunity (as with the modified live vaccine strains). This contrasts with highly virulent strains that cause acute symptoms including high fever, haemorrhages and central nervous system disorders, which often lead to high mortality close to 100%. Moderately virulent strains have variable disease outcomes; some pigs develop severe or chronic disease, whereas others fully recover. Some strains display distinct tropism for specific host tissues, which seem to be determined by the interaction of viral surface structures with cellular receptors exposed on the host cells. This seems to be the case for the highly virulent CSFV strain "Koslov" (Mittelholzer et al. 2000; Fahnøe et al. 2014c) that induces pronounced convulsions and seizures due to infection of cells in the central nervous system, thus showing an altered tropism compared to strains with lower virulence.

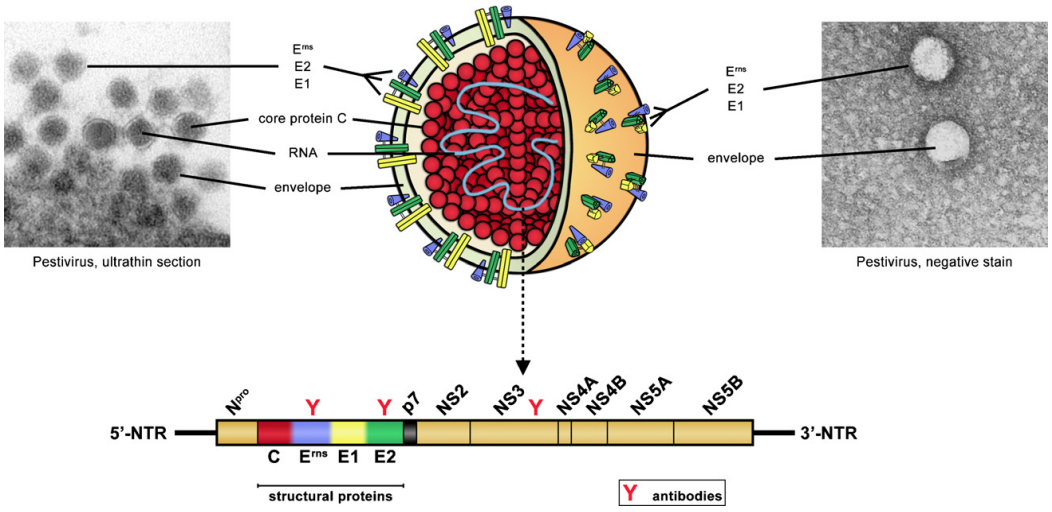


Figure 1. The CSFV particle is enveloped and contains a single stranded positive sense RNA genome of about 12.3 kilobases encoding a single large polyprotein. This polyprotein is cleaved by cellular and viral proteases into four structural proteins forming the CSFV particle (nucleocapsid protein C and the envelope glycoproteins E^{rns}, E1 and E2) and eight non-structural proteins (N^{pro}, p7, NS2, NS3, NS4A, NS4B, NS5A, NS5B) involved in virus replication, from Beer et al. (2007)

The virus has a positive sense single stranded RNA genome and is depicted in figure 1. The RNA is not capped like normal messenger RNA, but has an internal ribosomal entry site (IRES) in 5' NTR (non-translated region) or UTR (untranslated region). This IRES is an RNA structure responsible for binding the ribosomal translation initiation complex including the eIF3 translation initiation factor (Hashem et al. 2013). The genome is translated into one polyprotein that is subsequently cleaved by viral and host proteases into the mature viral proteins, which are divided into the structural proteins and non-structural. The virion consists of an envelope with the structural glycoproteins (E1, E2 and E^{rns}) embedded. The inside of the envelope is coated with the Core protein (C) that contains the RNA genome. There are eight non-structural proteins (N^{pro}, P7, NS2, NS3, NS4A, NS4B, NS5A and NS5B) of which NS3, NS4A, NS4B, NS5A and NS5B are necessary for replication (Mittelholzer et al. 1997; Moser et al. 1999). N^{pro} is both an autoprotease and an inhibitor of the antiviral type 1 interferon α/β response by inducing IRF3 degradation (Gottipati et al. 2013; Gottipati et al. 2014). NS5B the RNA dependent RNA polymerase (RdRp) is the active synthesizing protein involved in RNA replication.

Viral lifecycle

The viral lifecycle can be kick started by transfection of viral RNA into cells, in culture, like PK-15 or SK6, which indicates that the RNA is infectious. This is a feature that we exploit when testing full-length cDNA clones. The initial binding of the CSFV particle to host cells is thought to happen by interactions between viral envelope proteins present on the virus surface (E^{rns}, E1 and E2) and unknown cellular receptors present on the cell surface. Several research groups have studied the role of the envelope proteins including E2 in the virulence of CSFV (Van Gennip et al. 2004; Risatti et al. 2006). A range of host proteins that directly interact with the E2 protein of CSFV or BVDV have recently been identified including proteins that acts as receptors for other types of viruses (Gladue et al. 2014). An important cellular receptor for BVDV attachment has been identified as the cell surface protein CD46, which binds to BVDV and subsequently promotes virus entry into the cell (Maurer et al. 2004). However, the receptor(s) for CSFV remains to be identified. The 3-dimensional (3D) structure of BVDV E2 protein has recently been determined (El Omari et al. 2013; Li et al. 2013), which is a large step towards identification of virus-cell receptor interactions within BVDV and related viruses including CSFV. The amino acid identity between the E2 protein of BVDV and CSFV is about 65%, hence it is anticipated that the overall 3D structure of the two proteins will be highly conserved. After cell entry, the RNA is translated into the proteins mentioned above. The replication complex consisting of the non-structural proteins (NS3, NS4A, NS4B, NS5A and NS5B) starts to make negative strand RNA copies, which will function as templates for generating new positive strand RNA genomes for packaging into virions. The structural proteins together with the positive sense RNA will be packaged and leave the cell by excretion by the golgi apparatus (Murray et al. 2008).

Virus population diversity

The virus protein RdRp is a vital part of the replication complex that generates the new viral genomes. Due to the lack of error correction and the error-prone nature of the RdRp, RNA viruses have the highest known mutation rates (Drake 1993). This can be seen as both an

advantage and a potential danger for the virus. The high mutation rate might aid an escape from a host immune response that will be beneficial for the virus. However, if for example the mutation rate is too high each replication cycle will incorporate several mutations dispersed randomly that can be detrimental or even lethal for the virus population. It has been shown that a reduced mutation rate had an attenuating effect on the virus (Zeng et al. 2014). So, natural selection must play a vital role in fine-tuning the error rate of the RdRp of a viral population.

The high mutation rate leads to intra population diversity of closely related but slightly different genomes or haplotypes. This is described as a cloud of viral genomes dispersed around the consensus sequence in sequence space. In principal, all possible mutations should occur at every position in the genome, but transitions seems to be more common than transversions (Acevedo et al. 2014). This means that if a mutation gives a higher fitness it is picked up by selection and fixed in the population. We have observed this in both *in vitro* (in cell culture) and *in vivo* (in pigs) (Rasmussen et al. 2013; Fahnøe et al. 2014c).

Another concept is the quasispecies theory that also builds on the high mutation rate and the cloud of closely related genomes under selection. However, the theory predicts that the population will be most stable at a lower and flatter fitness peak in the fitness landscape compared to a higher and narrower peak (Lauring and Andino 2010). This means that selection is working on the viral cloud rather than on just the fittest haplotype. We have data from an *in vitro* study of BDV strain "Gifhorn" that support this theory (unpublished data). In this study, we found different fitness for two populations of the BDV "Gifhorn strain" with the same consensus sequence in a host adaptation study in primary pig kidney cells. The fittest population was derived from a cDNA clone and survived all ten passages. However, the other population died out after five passages and was the parental isolate from which the cDNA clone was derived. The experiments indicated that the structure of the population was responsible for the difference in fitness. Deep sequencing analysis will hopefully give some answers to this interesting result. The consensus sequence seems important, but it probably only exists as a distribution rather than being exactly represented on individual genomes (Wright et al. 2011; Acevedo et al. 2014). Recent studies by Sakoda's group in Japan (Tamura et al. 2012) revealed specific sequence changes (within E2 and NS4B coding sequences) that can be related to a gain of virulence. Sakoda's group identified specific amino acid residues in E2 and NS4B that are critical for determining the virulence difference between the vaccine

strain “GPE-“ and the highly virulent CSFV strain “ALD”. Importantly, modification of these residues in the two proteins had a synergistic effect on the virulence in pigs and on virus spread and replication efficiency in cell culture.

Molecular evolution

Molecular evolution is focused on molecular change in DNA, RNA or protein sequence over time to explain the adaptation of different organisms. This angle is interesting because all living organisms have evolved from some kind of common ancestor and should share some DNA. That DNA has changed due to mutation, which must explain some of the phenotypic differences between different species etc. Upon this constantly introduced variation by mutation acts random drift and selection that can be positive, neutral or negative.

RNA viruses such as CSFV have high mutation rates and evolve more rapidly than other organisms. Since functional genomes can transcomplement non-functional ones during replication the pedigree can become confusing and not as straightforward as for either pro- or eukaryotes. However, they must all have a functional ancestor within the phylogeny structure of the population. By using ancestral reconstruction that infers the sequence (DNA/RNA or protein) of internal nodes in a phylogeny the functional ancestor can be predicted. These sequences can be inferred and provide information about the evolution of different sites in the genome (Cai et al. 2004). Methods for ancestral reconstruction include parsimony (Williams et al. 2006), maximum likelihood, and Bayesian inference (Ronquist 2004). Bayesian inference has for example been used to reconstruct ancestral viral proteins in picorna viruses (Gullberg et al. 2010).

The variation created by mutation is subject to drift and selection within the viral population. Most mutations will either be neutral or deleterious and lead to neutral or negative selection, but some might improve fitness and lead to positive selection. Discerning sites and areas of the genome under positive selection is of great value because it provides information about adaptation and evolution patterns of viruses. One of the most common tools is the dN/dS ratio, where dN is the number of nonsynonymous mutations per nonsynonymous site and dS is the number of synonymous mutations per synonymous site, which describes the ratio of

nonsynonymous to synonymous substitutions. If the selection is neutral the dN/dS =1, if there is negative selection dN/dS < 1 and if there is positive selection dN/dS > 1. However, even though successful in identifying positive selection in HIV virus (Nielsen and Yang 1998), averaging whole genomes might be too conservative since most mutations are either neutral or detrimental. So different methods have been developed to look for positive selection in individual sites in tools like PAML (Yang 1997). Looking for dN/dS > 1 in viral NGS data has been tried (Morelli et al. 2013) but is still in its infancy because of issues in determining the population phylogeny.

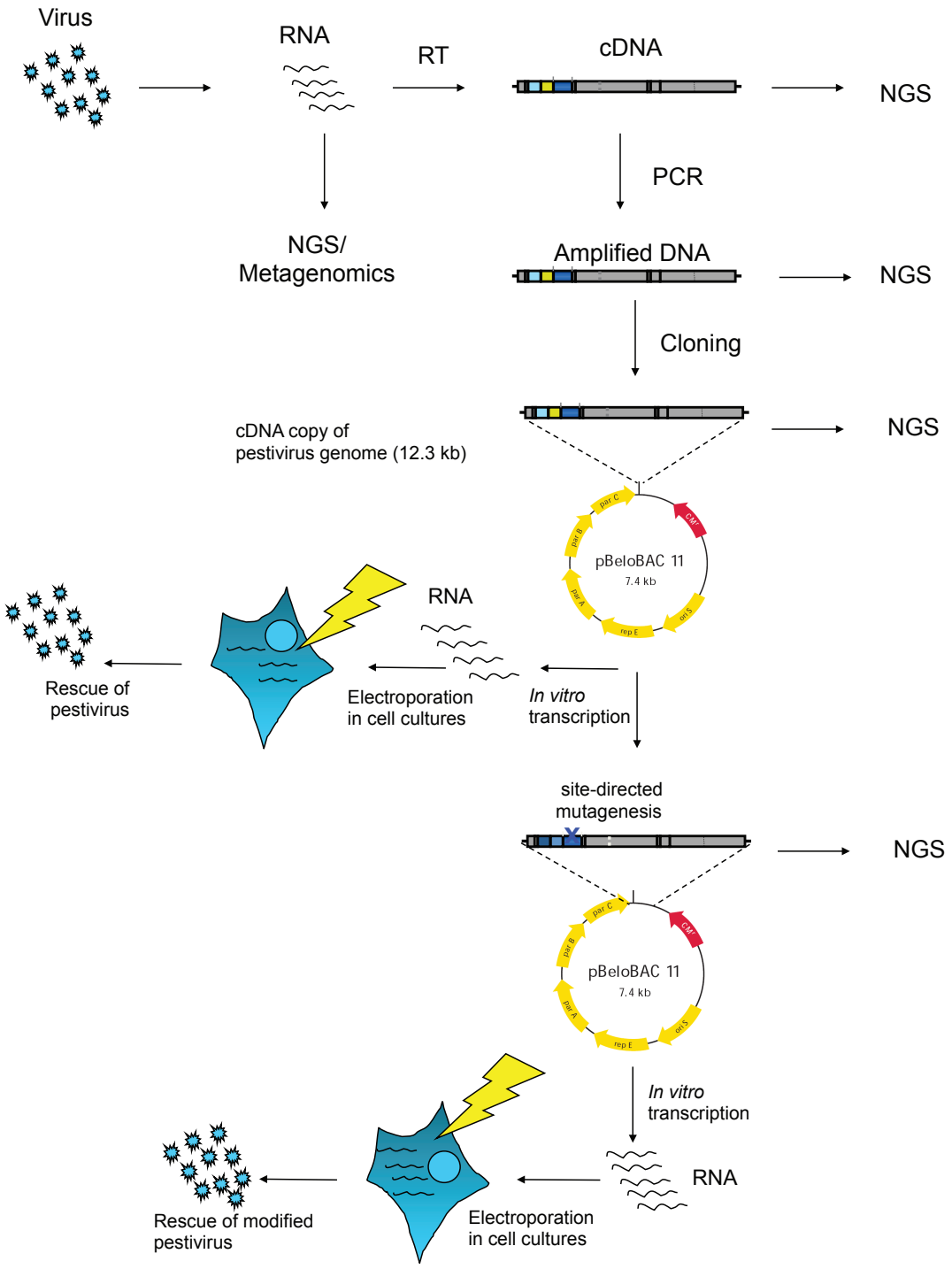


Figure 2. Pestivirus reverse genetics system. The cartoon shows the cloning procedure and the rescue of virus in cell-culture and also the NGS sequencing. In addition, the site-directed mutagenesis and rescue of modified virus is shown.

Reverse genetics

Reverse genetics is an approach where the phenotypical traits are investigated following sequence specific alterations. In order to establish reverse genetic systems, the virus cDNA genome must be cloned into a stable vector. This will allow subsequent mutagenesis and testing. The first functional CSFV cDNA clones obtained were from the CSFV strain "C" (Moormann et al. 1996) and the highly virulent CSFV strain "Alfort/187" (Ruggli et al. 1996). Other functional CSFV cDNA clones followed, for example the virulent strains "Alfort/Tübingen", "Eystrup" and "Brescia" (Meyers et al. 1996; Mayer et al. 2003; Van Gennip et al. 2004). In our lab we have developed a reverse genetic system for CSFV and BDV. The system is built on full-length cDNA clones inserted into a bacterial artificial chromosome (BAC) (Rasmussen et al. 2010) that ensures the stability of these large inserts (12.3 kb) (Rasmussen et al. 2013). Figure 2 describes how these full-length cDNA clones are generated by RT-PCR and in what way each step can be used for NGS sequencing. Once a full-length clone has been established, run-off RNA transcripts can be generated by in vitro transcription from an upstream T7 promoter. The RNA generated is potentially infectious and the infectivity is tested in cell culture by electroporation and subsequent immuno histochemical staining. This system allow for mutation of any nucleotide and subsequent potential rescue of virus in cell culture. The rescued modified viruses can then be used for animal infection experiments, as done in manuscript 3 and 4 or for *in vitro* studies. Two different mutation methods have been used with success, the targeted recombination-mediated mutagenesis using the Red/ET system (Rasmussen et al. 2013) that is precise but laborious. The second system is built on a modified site-directed mutagenesis approach (Risager et al. 2013) that is fast but not as accurate as the Red/ET approach. However, screening by Sanger sequencing and subsequent full-length sequencing by NGS has proven that out of two potential cDNA clones generated by the latter at least one will be a 100% match to the predicted sequence. Therefore the site-directed mutagenesis was chosen as the preferred technique, which has been used for manuscript 3 and 4 with success.

Next generation sequencing

Within the last 10 years the emergence of new sequencing technologies has altered the possibilities and the scope of applications within biological sciences. The old way that was used for sequencing the human genome for example was built on conventional Sanger sequencing and took almost ten years. Next generation sequencing (NGS) technologies allows a human genome to be sequenced in a single day for only a tiny fraction of the cost of the human genome project. The NGS technologies differ quite substantially in their chemistries, but have the PCR amplification step in common in the library preparation, and allow single DNA molecules to be sequenced resulting in multiple reads. In addition, no specific primers are used in the sequencing process, which is mainly an advantage. The read length varies from 40 bases to 1000 bases depending on the technology; in general a longer read is almost always preferred. It is the enormous number of reads per library that makes NGS so revolutionary compared to Sanger. An Illumina Miseq run can produce up to 50 million reads of 300 bases each which amounts to a total of 15 Gigabases (http://systems.illumina.com/systems/miseq/performance_specifications.html) compared to a typical Sanger read of 800 bases. These numbers vary between NGS platforms and are under constant development towards higher outputs and a longer read. A third generation of platforms are being launched that can produce extremely long reads (1000- 10.000) bases and does not require PCR amplification at all. However, these technologies are still in their infancies, the error rate is still too high and the input amount and purity of DNA is limiting but the potential of these platforms cannot be denied.

RNA virus applications of NGS

NGS has also made quite an impact on viral research. Because of the non-specific sequencing the technology can be used for pathogen discovery by metagenomics. This was used for the Schmallenberg virus, where this previously unknown orthobunyavirus was identified from only 7 reads out of more than ca. 30.000 reads in the sample library (Hoffmann et al. 2012). Metagenomics is a bioinformatic discipline that involves heavy computation and analysis.

Some of the challenges of this approach are found in the sample preparation, where host nucleotides can contaminate the sample and prevent the detection of the pathogen. In addition, the choice of sample to sequence in which the pathogen is present is also important. However, the potential and results of metagenomics will only strengthen the demands in the future. We have changed from Sanger to NGS for full-length sequencing of our cDNA clones and now also use NGS for full-length sequencing of different pestivirus isolates (Fahnøe et al. 2014a; Fahnøe et al. 2014b). This should at least provide a reliable consensus sequence. In addition, because NGS reads represent individual cDNA molecules it has the potential of deep sequencing to understand the underlying virus population diversity. Several studies have been performed in this area (Wright et al. 2011; Acevedo et al. 2014; Fahnøe et al. 2014c). Data analysis includes SNP analysis, haplotype reconstruction and molecular evolution. For CSFV, studies have shown that high heterogeneity is linked to high virulence (Töpfer et al. 2013). Other potential areas are RNA-seq of host mRNA to look for gene expression patterns and RNA structure analysis by NGS to mention a few.

NGS methods and results

Sample preparation for NGS

In order to get good results from NGS, sample preparation is crucial. If not carried out properly the library preparation will either fail or the sequencing result will not be of any use because of low quality sequence and few reads aligning to the reference sequence. Major issues are template amount, purity and integrity of the template, host RNA or DNA contamination and introduction of errors during sample preparation and sequencing. We have developed several methods and tested them (Fig. 2).

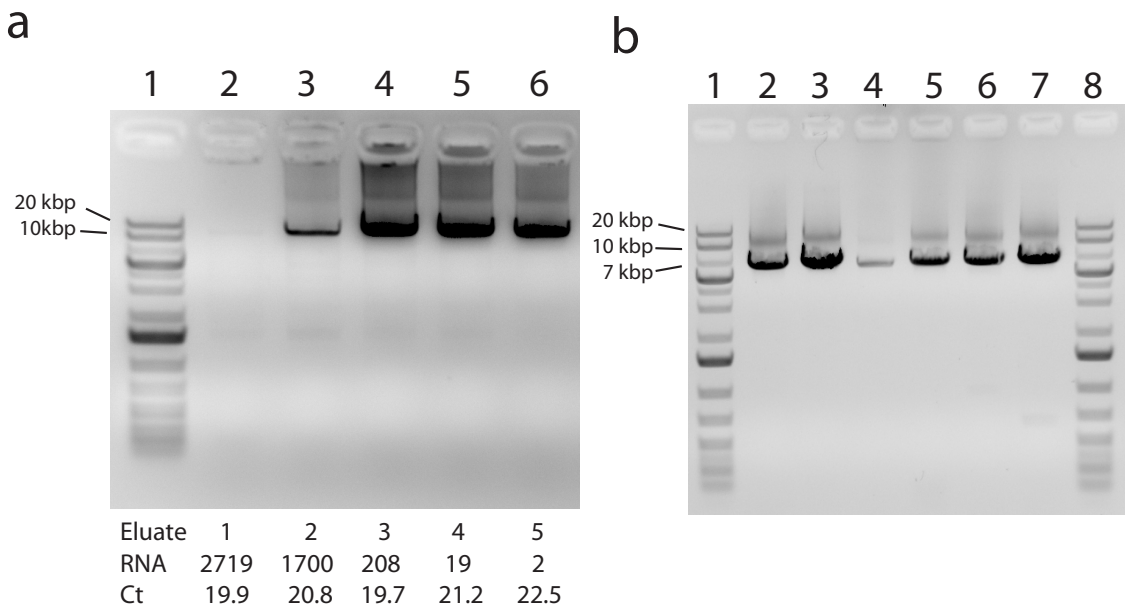


Figure 3. RT-PCR amplification. a) Full-length RT-PCR of tonsil RNA run on a 1% agarose gel. Underneath are described each eluate, total RNA concentration measured by nanodrop and viral RNA concentration in ct values measured by RT-qPCR. b) Half-length RT-PCR performed on 3 vaccinated pig serum samples visualised on a 1% agarose gel. Lane 1 and 8 depicts +1 kb marker; lanes 2, 4 and 6 are the 5' end fragment and lane 3, 5 and 7 are the 3' end fragment.

Full-length RT-PCR

Sequencing of CSFV from full-length RT-PCR products has so far seemed the most robust way of generating templates from high titer samples for library preparation (Leifer et al. 2010; Rasmussen et al. 2013; Fahnøe et al. 2014a). The method can be modified for different genotypes of CSFV by altering the 5′ primer, but because of the conserved 3′ end of the genome no primer alterations are needed in this end. This approach has also worked in sequencing of BDV (Fahnøe et al. 2014b). The RNA extraction procedure yields five consecutive eluates with a decreasing total amount of RNA. However, the yield of RT-PCR products from each eluate seemed to depend on the sample type. This was clearly seen for the tonsil samples where the proportion of full-length viral RNA compared to the total RNA was important for the success of the RT-PCR (Fig. 3a). A similar pattern was observed for whole blood samples (data not shown). Serum and cell culture samples had the strongest RT-PCR in the 1st eluate and decreasing tendency for each consecutive eluate (Fig. 4a and b). Taken together our results show that for a high complexity sample (tissue or blood) the 3rd to 5th eluate gave the strongest RT-PCR product while for low complexity samples (serum or cell culture) the 1st or the 2nd eluate performed the best. We were consistently able to obtain full-length RT-PCR products from high titer samples (Ct < 24).

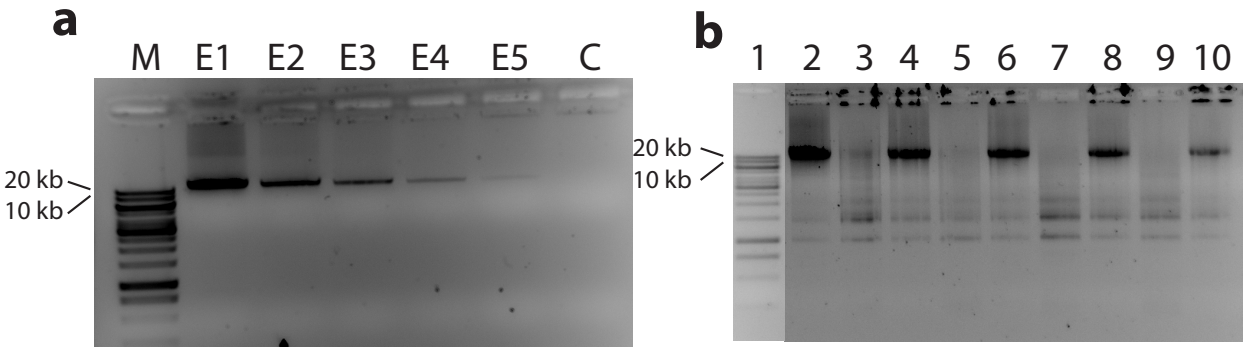


Figure 4. Cell culture and serum full-length RT-PCR. a) Full-length RT-PCR of cell-culture supernatant RNA run on a 1% agarose gel. Depicting M: +1kb and E1-E5: eluates 1-5 and C : negative control. b) Full-length RT-PCR of cell-culture supernatant RNA run on a 1% agarose gel. Lane 3, 5, 7, 9 are different eluates with SYBR green added to the mastermix. Lanes 2, 4, 6, 8 and 10 are eulates 1-5 respectively without SYBR green added to the mastermix. Lane 1 is +1kb.

Half-length RT-PCR

As mentioned above, low titer samples are challenging. Therefore, we developed a robust method to amplify the genome in two overlapping fragments. This method allowed samples with Ct values up to 32 to be amplified and deep sequenced. The method is described in manuscript 5 and has proven successful for nasal swabs, serum and blood samples (fig. 3b). Preliminary results have shown that it also works on tissue samples (data not shown). The two fragments can easily be pooled in equal amounts and thereby save a library preparation. This has, so far, proven to be the most successful method in our hands to address the low titer samples.

RNA sequencing

We had the opportunity to try RNA sequencing on the FLX platform with library preparation performed by our collaborators at FLI Germany. The Trizol RNeasy protocol allowed the extraction of highly pure and high integrity RNA giving 5 eluates. The samples chosen were three CSFV serum samples described in manuscript 5, one blood sample (eluate 3 and 4) and a BDV "Gifhorn" virus from a cell lysate (SFT-R cells). 200 ng RNA of each sample was sent to FLI for NGS sequencing. Figure 5 shows the distribution of reads mapped to each respective reference sequence. Serum samples have mapping percentages close to 100%, the blood about 72%, while the BDV cell culture sample is below 1%. Metagenomic analysis confirmed the majority of the rest of the reads to belong to either pig (Sus scrofa) for the CSFV samples and sheep (Ovis aries) for BDV sample (data not shown). So the high titer blood and serum samples gave excellent results while the BDV sample was highly contaminated by host RNA probably due to the freezing and thawing of the culture leading to cell rupture before RNA extraction. However, because of the high number of reads in the RNA sequencing the entire genome was covered except for the extreme ends leading to a consensus sequence identical to the NGS of a full-length RT-PCR product obtained from the same RNA (Fahnøe et al. 2014b). RNA sequencing worked well for the high titer samples from blood and serum. However, this kind of sequencing is highly sensitive to contamination from host or other sources.

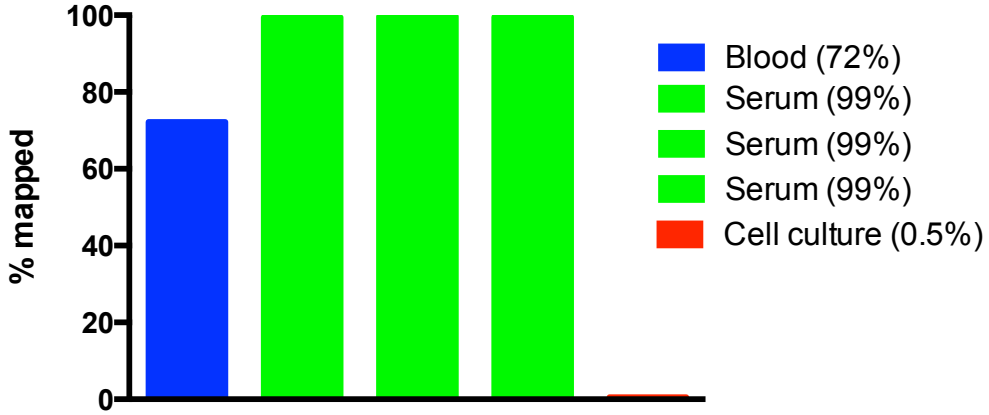


Figure 5. RNA sequencing. Depicts the fraction of reads mapped to the reference genome in %.

Second-strand cDNA sequencing

Most NGS library preparation kits do not take single stranded DNA, which makes second strand cDNA synthesis necessary. We have only done preliminary experiments into this approach. However, this approach has the advantage over the RNA sequencing because the cDNA is more stable and it can more easily be send to outside laboratories for library preparation. This is because the cDNA is not considered infectious material after the removal of RNA. Two different first-strand priming approaches were tried out as tests. The first was random hexamers (RH), which should target all RNA in the sample thereby being unspecific, while the second was a double specific primed RT (2XRT) as described in manuscript 5. The sample chosen was a first eluate taken from an extraction of a cell culture supernatant vRos_P1 ($log_{10}TCID_{50}$ = 6.8/ml), described in manuscript 4, this time before freezing. Superscript III was used for 1st strand synthesis for both and the maximum RNA volume of 8 µl was applied. Two different temperature profiles were chosen to accommodate the different priming methods (RH: 25 °C, 15 min; 50 °C, 90 min; 85°C, 5 min. 2XRT: 50°C, 90 min; 85°C, 5 min). After first strand synthesis the NEBNext® Second Strand Synthesis Module kit (New England Biolabs) was used. The kit involves RNase H, DNA Polymerase I and a DNA Ligase. In addition, it is completely compatible with Superscript III and involves only a few manipulations. After second-strand synthesis the product was purified on a PCR spin column and sent for library preparation. The results revealed two different results with interesting sequence depth profiles (Fig. 6a and b). A close to uniform depth was observed for the RH with the depth going to zero at the extreme ends, while the 2XRT had high coverage in the 3' end corresponding to the 3' primer then a gradual decrease until position 6000 where a small increase was seen associated with the second primer position and then a further decrease towards zero until the 5' end. A surprising result was found in the fraction of reads mapped to the reference genome (Fig. 6c). 2% of the reads aligned to CSFV for the RH that could be explained by the rest being host specific. However, the 2XRT had only 1.5% reads mapped; this was a surprise because two specific primers only primed the RT. Further metagenomic analysis must be performed to determine what the rest of the reads represent. So a combination of the two methods might be the most optimal protocol with DNase and/or RNase treatment to remove host DNA and RNA as described by others (Logan et al. 2014). The

potential of this protocol, besides removing the PCR amplification, is that it is easily applicable to other RNA viruses.

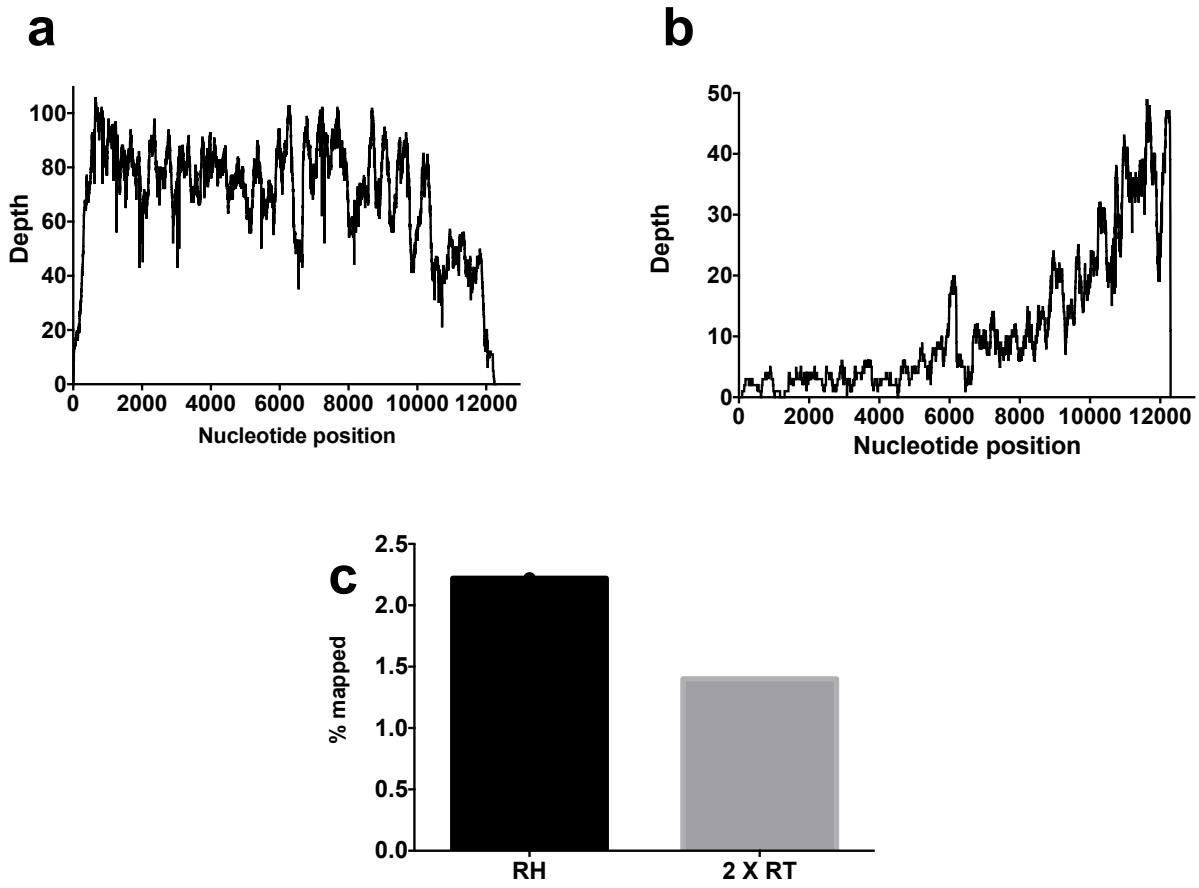


Figure 6. Second-strand cDNA sequencing. a) Coverage plot depicting sequence depth of 1st strand synthesis primed by random hexamers (RH). b) Coverage plot depicting sequence depth of 1st strand synthesis primed by 2 X specific primers (2XRT). c) Depicts the fraction of reads mapped to the reference genome in %.

Library preparation

None of our samples were sequenced by NGS in-house. Either Dr. Dirk Höper's group at FLI Insel Riems, Germany sequenced the samples using the FLX or the Illumina Miseq, or they were sent to the DMAC (DTU multi-core assay center) within DTU Systems Biology for Ion PGM sequencing. This approach meant we were not personally involved in handling the

library preparation but were able to discuss kits and procedures. It also resulted in a consistent quality of data of the more than 400 samples sequenced and analyzed during my PhD. In addition, it allowed me to apply my time elsewhere.

NGS data analysis and pipeline

We have set up a data analysis automated pipeline for RNA viruses that outputs plots generated in R along with consensus FASTA, SNP call, sequence depth and mapping (Fig. 7). The whole pipeline is a combination of bioinformatics tools described below and set up as a Python script. The only things you need are to choose the reference (from a list of predefined reference sequences in a reference database), input raw FASTQ file and trimming parameters for each sequenced library. The pipeline was created to accommodate the most frequent NGS analysis requests for viral samples and has been tested for foot-and-mouth disease virus (FMDV), BDV, Aleutian disease virus (ADV) in addition to CSFV and could easily be adapted to other RNA viruses. Below each paragraph are shown the command line used and various R and python scripts developed. The tools were installed in a UNIX shell and run from the command line. This approach has several advantages compared to graphic all-in-one tools. The UNIX shell allows for coupling of different tools and immediate data manipulation by custom-made scripts. New bioinformatics tools are usually designed for UNIX and if maintained updated regularly, which makes them superior to the all-in-one graphic application that cannot be updated on every type of analysis. Analysis of RNA virus deep sequenced samples also possesses challenges that the eukaryotic tools were not designed to handle as described below.

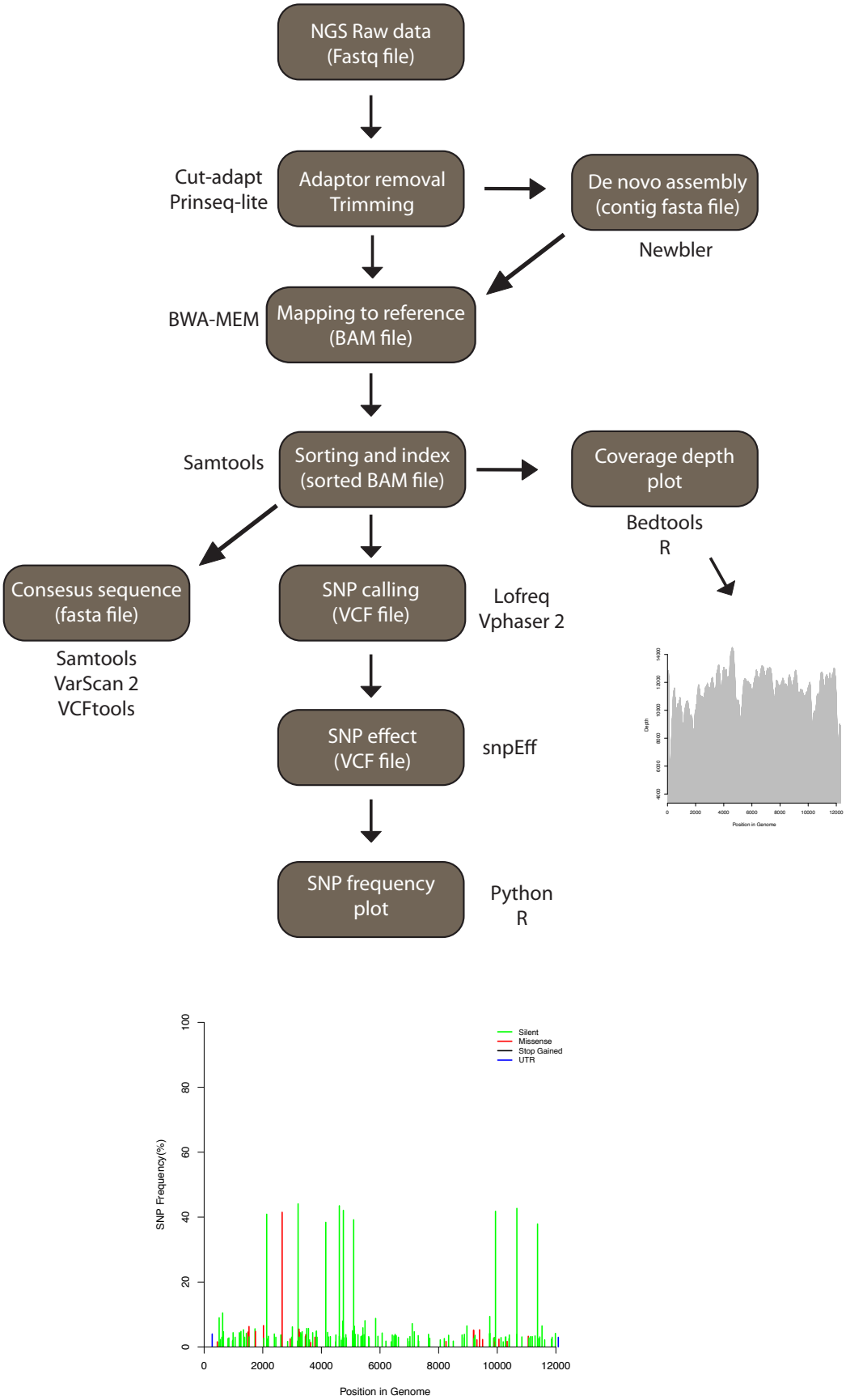


Figure 7. NGS data analysis pipeline. The figure depicts the data analysis workflow and outputs made by different tools in the pipeline. The tools are mentioned outside each box and inside there are descriptions of the process and the file format.

Raw data: Fastq or SFF file

Raw data was received either as a FASTQ or an SFF file. The SFF (standard flowgram format) file is a binary file, which includes the flowgram information from the FLX pyrosequencing. The SFF file includes more information compared to the FASTQ file that basically includes the sequence and quality score of each read. However, so far the Newbler software (Roche) is the only mapper or assembler that uses this extra information. This does not allow for read quality trimming and filtering before mapping or assembly. So in most cases we prefer to convert the SFF file to FASTQ format performed by sffinfo (Roche) combined with a python script called convertFastaQualtoFastq.py because the output from sffinfo is a fasta and qual file, which needs to be combined.

Command line:

#Convert ssf file to fasta and qual

sffinfo -s (sff file) > *.fasta

sffinfo -q (sff file) > *.qual

#Convert fasta and qual to fastq

convertFastaQualtoFastq.py --fasta *.fasta --qual *.qual

Read Quality assessment and trimming and filtering

The FastQC tool (Andrews 2010) that can both be used as a graphical application or as a command line tool analyzed the raw FASTQ file. The output is a HTML file that can be read in a standard browser. The analysis was used extensively and presented as a poster at the EPIZONE 2013 (Title: Comparison of two Next Generation Sequencing platforms for full genome sequencing of Classical Swine Fever Virus) and by Rasmussen et al. (2013). In short, the output is a read quality analysis in respect to read length, mean quality, nucleotide overrepresentation, mean read length and sequence adapter over representation etc. Subsequently, the analysis provides the parameters for filtering and trimming of the reads. Cutadapt (Martin 2011) removes primer or adapter and the output that is piped into prinseq-lite (Schmieder and Edwards 2011) for filtering and trimming. The output is an edited FASTQ file ready for de novo assembly or mapping to a reference.

Command line:

 # cutadapt (can be piped into priseq-lite)

cutadapt -O (length of primer) -e (error fraction of primer permitted 0.05-0.1 for 20 bp primer) -a primer3 -g primer5 *.fastq > *.out.fastq

#Trimming by prinseq-lite output trimmed fastq example FLX

cat *.fastq | prinseq-lite.pl -fastq stdin -out_good stdout -out_bad null -trim_to_len 600 -trim_left 2 trim_qual_left 20 -trim_qual_right 20 -min_len 100 -min_qual_mean 20 > *.trim.fastq

NGS error correction and SNP call benchmarking

FLX and Ion PGM data are prone to indel errors caused by homopolymers. In order to address this problem, a full-length PCR product and a corresponding full-length RT-PCR product were sequenced by the Ion PGM. The first product was obtained from a defined cDNA clone (Kos_ELP1) representing a uniform clonal population. The second product represents a virus population and was obtained from a RT-PCR product (vKos_ELP1/P-1) derived from virus

rescued after electroporation of run-off transcripts in PK-15 cells followed by one cell culture passage.

Several tools for error correction were applied to these two NGS data sets, including Coral (Salmela and Schroder 2011),

InDelFixer (http://www.cbg.ethz.ch/software/InDelFixer/) and RC454 (Henn et al. 2012). After error correction by each tool, mapping of reads was performed using BWA MEM (Li 2013), Mosaik (Lee et al. 2014) or Bowtie 2 (Langmead and Salzberg 2012) followed by SNP calling performed by V-phaser 2 (Yang et al. 2013), Lofreq (Wilm et al. 2012) or VarScan 2 (Koboldt et al. 2012). By comparing the clonal and the virus NGS data sets it was possible to benchmark the different error correction tools. RC454 significantly outperformed both Coral and InDelFixer by removing all detectable indels above 1% detected by VarScan 2 from the clonal sample (Fig. 8) and reducing the SNPs called above 1%. RC454 did not remove the SNP variation in the viral sample but increased the amount of SNPs detected (Fig. 8b) and only removed the indels called. A similar result could be observed when V-phaser 2 called the SNPs and indels with or without RC454 (Fig. 8c and 8d).

Command line:

#RC454 1

*perl runMosaik2.pl -fa *.fasta -qual *.qual –ref reference.fasta -o * -qlx -param454*

#RC454 2

*perl rc454.pl *.qlx *.fasta *.qual reference.fasta * -RC454 –bam*

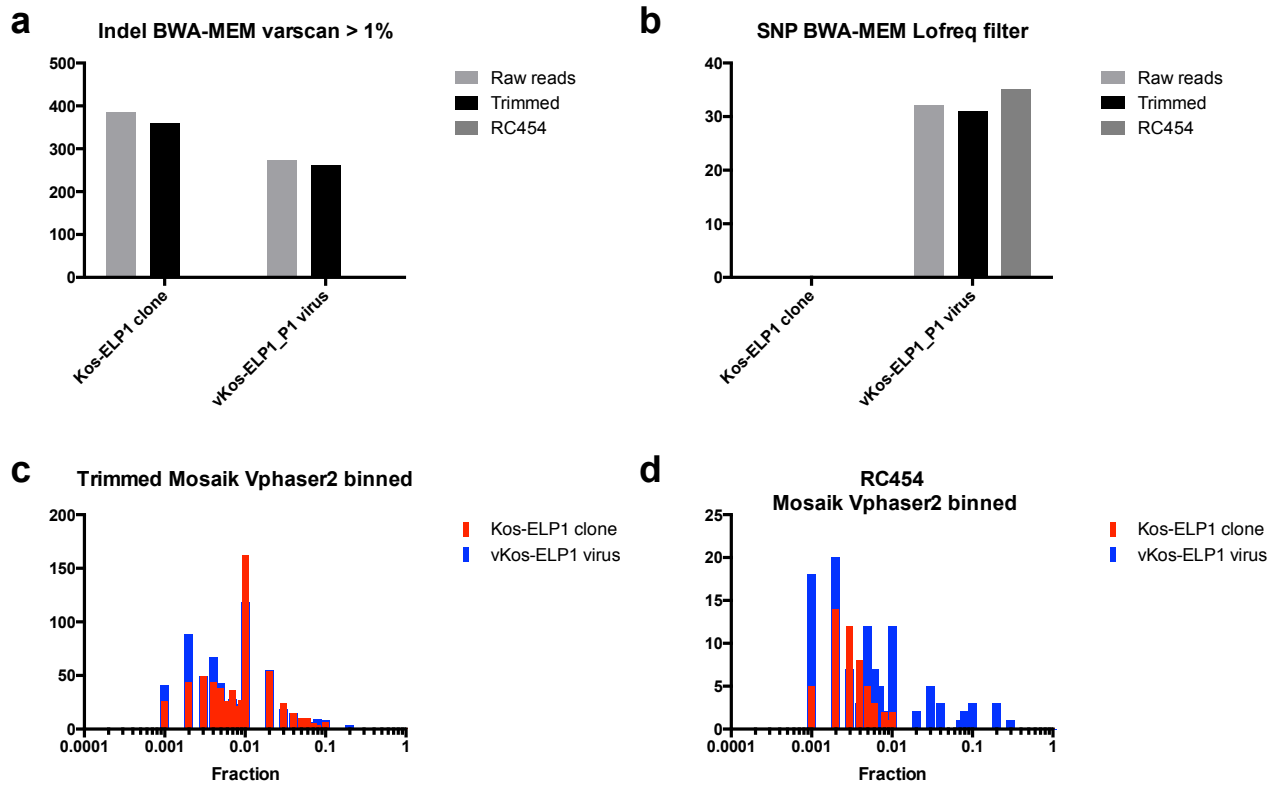


Figure 8. NGS data error correction. a) Indels detected above 1% by VarScan 2. b) SNPs detected by lofreq. c) SNPs called by V-Phaser 2 after read trimming. d) SNPs called by V-Phaser 2 after RC454 error correction.

De novo assembly

The de novo assembly does not take any sample information into account, but simply tries to assemble the reads into long contigs. This method of generating a consensus sequence does not require a reference sequence and was actually used to generate references for subsequent mapping (Fahnøe et al. 2014a). We used the Newbler software for de novo assembly of FLX and Ion PGM data. The assembler can take input as SFF, FASTQ or FASTA files. The output is a project folder containing run statistics and files containing the contig data. Most times we were able to retrieve one large contig spanning the entire CSFV genome (12.3 kb). However, sometimes ultra-deep sequencing with above 1000X coverage gave rise to unsuccessful assemblies containing multiple contigs. A solution could be to downsize the number of reads so the depth was reduced to approx. 100X. This can be done by seqtk that takes a fraction or a specific number of reads from a FASTQ file. In most cases, that had a beneficial effect. De novo assemblies were used to generate consensus and reference sequences for several studies

(Risager et al. 2013; Rasmussen et al. 2013; Kvisgaard et al. 2013; Fahnøe et al. 2014a; Fahnøe et al. 2014b).

Command line:

#De novo assembly applying Newbler

newAssembly (project name)

addRun -lib wgs (project name) *.trim.fastq

runProject -cpu 2 (project name)

#Get a subset of reads from a fastq file (can help to get longer contigs)

seqtk sample *.fastq [ratio float or number of reads] > *.sub.fastq

Mapping reads to a reference sequence

As mentioned above a suitable reference sequence is needed to obtain decent mapping results. This means that many mismatches between the sample consensus and the reference leads to poor alignments of the reads and many reads not being mapped. So the reference should represent the sample as closely as possible. If the sequence is not known a de novo assembly is recommended. However, if many samples with slight consensus differences are to be compared then the same reference must be used. Mapping the reads to the sample consensus will give the intra sample comparison instead. We have tested three different mappers BWA-MEM (Li 2013), Bowtie 2 (Langmead and Salzberg 2012) and Mosaik (Lee et al. 2014). Overall, they all performed well with close to 100% of the reads mapped to the reference. However, a slight difference was observed when the extreme ends were compared. Bowtie 2 seemed to have difficulties aligning reads to 5' and 3' ends resulting in poor coverage. BWA-MEM uses local alignment that is designed for reads longer than 70 bp and

performed fast and accurate. The output format is SAM (sequence aligment map), which is TAB-delimited text format. This format is normally immediately converted to its binary BAM format (binary alignment map) because it takes up less space and most downstream applications run on the BAM format. SAMtools (Li et al. 2009) is a suite of tools designed for the manipulation of the SAM/BAM format. The BAM file is both sorted and indexed by SAMtools, which is necessary for downstream application and visualization.

Command line:

#Mapping to Reference genome

bwa index *.fasta

bwa bwasw *.fasta *.trim.fastq | samtools view -Sb - > *.trim.bam

bwa mem *.fasta *.trim.fastq | samtools view -Sb - > *.trim.bam

#mapping with Bowtie 2

bowtie2-build *.fasta (reference name)

bowtie2 -x (reference name) -U *.trim.fastq | samtools view -Sb - > *.trim.bowtie2.bam

#mapping with mosaic

perl ~/NGS_Tools/RC454_SoftwarePackage/runMosaik2.pl -fa FASTA.file -qual QUAL.file -ref REF.file -o OUTPUTNAME -qlx -param454

#sort Bam file

samtools sort *.trim.bam *.trim.sort

#index Bam file

samtools index *.trim.sort.bam

Extraction of consensus sequence and coverage depth

After mapping and sorting of the BAM file the consensus sequence can be extracted. Both methods described involve SAMtools and SNP calling. SAMtools has a module that makes a pileup of overlapping reads for each position which can be used for SNP calling. The pileup is used to generate a BCF file, which is the binary format of the VCF file (variant call format). The VCF file can then be interpreted using VCFtools (Danecek et al. 2011) and a new consensus generated as a FASTQ file. However, the problem with this approach is that it cannot detect indels (insertions or deletions) in the consensus. So a modified version of SNP calling and generation of consensus was devised. The new approach involved a SAMtools pileup subsequent SNP and indels called by VarScan 2 (Koboldt et al. 2012) and a consensus sequence obtained from VCFtools. This approach was also able to incorporate indels in the new consensus sequence, which led to this being the choice of the pipeline. Coverage depth can be calculated by several tools and plotted. We chose the BEDTools (Quinlan and Hall 2010), which outputs a list of depth at each position that is plotted in R.

Command line:

#Getting consensus sequence using samtools in fastq format

samtools mpileup -uf reference.fa *.bam | bcftools view -cg - | vcfutils.pl vcf2fq > *.fq

#calling Indels and SNPs with VarScan and samtools mpileup and gunzip

samtools mpileup -B -f reference.fasta *.trim.sort.bam | java -jar VarScan.v2.3.6.jar mpileup2cns --min-var-freq .5 --output-vcf 1 | bgzip > *.con.vcf.gz

```
#indexing vcf file with tabix

tabix -p vcf  *.con.vcf.gz"


#get consensus with Vcftools + renaming fasta header to prefix

cat reference.fasta| vcf-consensus *.con.vcf.gz | sed "s/genome/ */" > *.fasta


#Create coverage by each position in the genome (Note the -d)

bedtools genomecov -d -ibam *.sort.bam > *.trim.sort.bam.cov


#Setting parameters for coverage plot in R, can be run as a R-script

test <- read.table('*.trim.sort.bam.cov')

pdf('*-cov.pdf')

plot(V3 ~ V2, xlim=c(0,12350), data=test, type="h", xlab="Position in genome", ylab="Depth",
lwd=0.25, yaxs="i", xaxs="i", frame=F, col=c("grey"), main="Title")

dev.off()
```

SNP calling and SNP translation

Normally, SNP calling tools are designed for eukaryotes where the frequencies follow homozygous/heterozygous diploid pattern. However, RNA viruses are known to generate large numbers of mutations and the population within a sample is heterogeneous. Several tools have been developed to look for low frequent SNPs and basically they filter the raw output and output a VCF file or a text file with variant positions and frequencies. We tested V-Phaser2 (Yang et al. 2013) and Lofreq (Wilm et al. 2012). Both the Mosaik and Bowtie 2 aligner was used and bowtie2 seemed to have difficulties at both ends of the genome (data

not shown). V-Phaser 2 could not process BWA MEM mapped bam files because of incompatibility. The results clearly showed that without error correction with RC454 real SNPs and indels could not be discerned from errors using V-Phaser 2 (Fig. 8c and d). Still V-Phaser 2 calls SNPs and indels in the clonal sample. However, after RC454 a lower threshold between 0,5 and 1% could be established for V-Phaser 2. There is a limit to how deep into the low frequency SNPs that can be trusted. This is defined by the error rates of the sequencing platforms and errors introduced during sample preparation. In manuscript 5 we tested this and the cut-off seemed to be between 0.5-1% for Ion PGM data. Lofreq was more stringent when compared to V-Phaser 2 with no SNPs detected in the clonal sample and same SNPs at equal frequencies above 1% in the viral sample as V-Phaser 2 (data not shown). However, V-phaser 2 can complement that Lofreq is not able to call indels.

We looked for a way to automate the translational interpretation of the SNP VCF files and found a tool called snpEff (Cingolani et al. 2012) that was designed for the Drosophila genome. In short, if the genome is not in the reference database a new reference has to be built. Then the VCF file can be processed and the translational effects added to a new VCF file. The output is not TAB-delimited and therefore not easily plotted. Therefore, we developed a python script to split the VCF file into TAB-delimited text output, which was plotted in R.

Command line and scripts:

#SNP calling with lofreq

lofreq_snpcaller.py -f reference.fasta -b *.trim.sort.bam -o *.vcf --format vcf --dont-join-mapq-and-baseq

#SNP effect using snpEff

#Create library add following lines to snpEffect.config for more see http://snpeff.sourceforge.net/supportNewGenome.html

CSFV genome, [name]

*.genome : [Name]

#build library

java -jar snpEff.jar build -genbank -v [name]

#SNP effect with snpEff

java -Xmx4g -jar ~/snpEff/snpEff.jar -c ~/snpEff/snpEff.config -v genome *.vcf > *.snpeff.vcf

```
#Python script to split VCF snpEff file
mydata = open(inputfile).readlines()
outfile = open(outputfile, "w")
for line in mydata:
  if line.startswith("##"):
    pass
  elif line.startswith("#"):
    line.replace("#", "")
    headerwords = line.split()
    outline = "{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\n".format(headerwords[0].replace("#", ""), headerwords[1], headerwords[3], headerwords[4], headerwords[5], "AF", "DP", "ref-forward", "ref-reverse", "alt-forward", "alt-reverse", "SB", "Effect", "Functional_Class", "Codon_Change", "Amino_Acid_Change")
    outfile.write(outline)
```

```
        else:
            words = line.split()
            semmicolonwords = words[7].split(";")
            pipewords = semmicolonwords[4].split("|")
            for i in range(len(pipewords)):
                if pipewords[i] == "":
                    pipewords[i] = "-"

            strandwords = semmicolonwords[2].split(",")
            outline = "{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\t{}\n".format(words[0], words[1], words[3], words[4], words[5], semmicolonwords[0].replace("AF=",""), semmicolonwords[1].replace("DP=",""), strandwords[0].replace("DP4=", ""), strandwords[1], strandwords[2], strandwords[3], semmicolonwords[3].replace("SB=",""), pipewords[0].replace("EFF=", ""), pipewords[1], pipewords[2], pipewords[3])
            outfile.write(outline)

    outfile.close()
    #R-script for plotting SNP
    test <- read.table('*.snpeff.txt', header=TRUE)"
    pdf('*-SNP.pdf')
    plot(100*AF ~ POS, col=ifelse(Functional_Class=="NONSENSE", "black", ifelse(Functional_Class=="MISSENSE", "red", ifelse(Functional_Class=="-", "blue", "green"))),
```

ylim=c(0,100), xlim=c(0,12350), data=test, type="h", xlab="Position in Genome", ylab="SNP Frequency(%)", main="Title", lwd=2, yaxs="i", xaxs="i", frame=F)

legend("topright", c("Silent", "NON-Silent", "Stop Gained", "UTR"), lty=1, lwd=2, col=c("green", "red", "black", "blue"), bty="n", cex=.75)

dev.off()

#Vphaser 2 SNP calling

variant_caller -i Kos-ELP.1-RC454_final.sort.bam -o ./

Manuscripts

Manuscript 1

Complete genome sequence of border disease virus genotype 3 strain Gifhorn

genomeA
Journals.ASM.org

Complete Genome Sequence of Border Disease Virus Genotype 3 Strain Gifhorn

Ulrik Fahnøe,[a] Dirk Höper,[b] Horst Schirrmeier,[b] Martin Beer,[b] Thomas Bruun Rasmussen[a]

DTU National Veterinary Institute, Technical University of Denmark, Lindholm, Kalvehave, Denmark[a]; Institute of Diagnostic Virology, Friedrich-Loeffler-Institut, Greifswald-Insel Riems, Germany[b]

The complete genome sequence of the genotype 3 border disease virus strain Gifhorn has been determined; this strain was originally isolated from pigs. This represents the consensus sequence for the virus used to produce the bacterial artificial chromosome (BAC) cDNA clone pBeloGif3, which yields a virus that is severely attenuated in cell culture.

Received 13 December 2013 **Accepted** 19 December 2013 **Published** 16 January 2014

Citation Fahnøe U, Höper D, Schirrmeier H, Beer M, Rasmussen TB. 2014. Complete genome sequence of border disease virus genotype 3 strain Gifhorn. Genome Announc. 2(1):e01142-13. doi:10.1128/genomeA.01142-13.

Copyright © 2014 Fahnøe et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported license.

Address correspondence to Thomas Bruun Rasmussen, tbrur@vet.dtu.dk.

Border disease virus (BDV) belongs to the genus *Pestivirus*, which includes other important animal pathogens, such as bovine viral diarrhea virus (BVDV) and classical swine fever virus (CSFV). BDV causes disease mainly in small ruminants, but cattle, pigs, and other wildlife may also be infected; indeed, the BDV genotype 3 (BDV-3) strain Gifhorn was originally isolated from both infected pigs and sheep kept on the same farm (1, 2). The BDV genome consists of a positive-sense RNA approximately 12.3 kb in length, which encodes a single polyprotein that is posttranslationally cleaved to form structural and nonstructural proteins. The genetic diversity of BDV is high, with multiple major genotypes reported (3). Complete genomic sequences have been described for genotypes BDV-1 (strains BD31 [4] and X818 [5]), BDV-2 (strain Reindeer-1 [6]), and BDV-4 (strain H2121 [3]). In addition, recently, the complete genome of BDV JSLS12-01, which is closely related to that of BDV-3 Gifhorn, has been reported (7). A complete sequence for BDV-3 Gifhorn was published previously, but this was derived from a bacterial artificial chromosome (BAC) cDNA clone, pBeloGif3 (GenBank accession no. GQ902940); virus rescued from this cDNA displays severe growth attenuation in cell culture (8). The complete genome sequence of the BDV-3 Gifhorn isolate has now been determined.

Viral RNA was extracted from BDV-3 Gifhorn-infected sheep fetal thymoid (SFT-R) cells, and full-length viral cDNAs were amplified by long reverse transcription-PCR (RT-PCR) as previously described (8). cDNA was prepared from viral RNA according to the manufacturer's protocol (Roche, Mannheim, Germany; cDNA Rapid Library Preparation Materials and Methods Manual). Sequencing libraries were generated from cDNA or from RT-PCR products using the SPRIworks Fragment Library System II (Beckman Coulter, Krefeld, Germany) and were sequenced using a 454 FLX (Roche). Newbler (Roche) was used for *de novo* assembly and for mapping of the reads using pBeloGif3 (GenBank accession no. GQ902940) as a reference sequence. Finally, the consensus sequences were aligned using MAFFT in the Geneious software platform (Biomatters).

The consensus sequence for the BDV-3 Gifhorn genome was obtained by replicate sequencing of RT-PCR products and RNA. The replicate samples generated the same consensus sequence, which was 12,325 nucleotides (nt) long. The polyprotein-coding sequence is 11,694 nt long and contains 3,898 codons. The 5′ and 3′ untranslated regions (UTRs) are 375 and 256 nt long, respectively. A comparison with the cloned Gifhorn genome sequence (from pBeloGif3, 12,326 nt) revealed a 1-nt deletion in the 5′ UTR and 11 nt differences in the coding sequence, including 8 that are nonsynonymous. The complete consensus sequence of this BDV-3 virus and its comparison with the pBelGif3 sequence should enhance our understanding of the factors that are important for viral attenuation. The generation of additional genomic data for BDV will assist further investigations into the properties of this group of viruses.

Nucleotide sequence accession number. The genome sequence of BDV-3 Gifhorn has been deposited in GenBank under the accession no. KF925348.

ACKNOWLEDGMENT

This work was supported by the Danish Research Council for Technology and Production Sciences (DRCTPS) (grant no. 274-07-0198).

REFERENCES

1. **Becher P, Avalos Ramirez R, Orlich M, Cedillo Rosales S, König M, Schweizer M, Stalder H, Schirrmeier H, Thiel HJ.** 2003. Genetic and antigenic characterization of novel pestivirus genotypes: implications for classification. Virology **311**:96–104. http://dx.doi.org/10.1016/S0042-6822(03)00192-2.
2. **Schirrmeier H, Strebelow G, Depner KR, Beer M.** 2002. Heterogenicity of pestiviruses: determination and characterisation of novel genotypes and species, p 32. Fifth Pestivirus Symp. Eur. Soc. Vet. Virol., Cambridge, England, 26 to 29 August 2002.
3. **Vilcek S, Willoughby K, Nettleton P, Becher P.** 2010. Complete genomic sequence of a border disease virus isolated from Pyrenean chamois. Virus Res. **152**:164–168. http://dx.doi.org/10.1016/j.virusres.2010.05.012.

4. **Ridpath JF, Bolin SR.** 1997. Comparison of the complete genomic sequence of the border disease virus, BD31, to other pestiviruses. Virus Res. **50:**237–243. http://dx.doi.org/10.1016/S0168-1702(97)00064-6.
5. **Becher P, Orlich M, Thiel HJ.** 1998. Complete genomic sequence of border disease virus, a pestivirus from sheep. J. Virol. **72:**5165–5173.
6. **Avalos-Ramirez R, Orlich M, Thiel HJ, Becher P.** 2001. Evidence for the presence of two novel pestivirus species. Virology **286:**456–465. http://dx.doi.org/10.1006/viro.2001.1001.
7. **Liu X, Mao L, Li W, Yang L, Zhang W, Wei J, Jiang J.** 2013. Genome sequence of border disease virus strain JSLS12-01, isolated from sheep in China. Genome Announc. **1**(6):e00502-13. http://dx.doi.org/10.1128/genomeA.00502-13.
8. **Rasmussen TB, Reimann I, Uttenthal A, Leifer I, Depner K, Schirrmeier H, Beer M.** 2010. Generation of recombinant pestiviruses using a full-genome amplification strategy. Vet. Microbiol. **142:**13–17. http://dx.doi.org/10.1016/j.vetmic.2009.09.037.

Manuscript 2

Complete genome sequence of classical swine fever virus genotype 2.2 strain Bergen

genomeA
Journals.ASM.org

Complete Genome Sequence of Classical Swine Fever Virus Genotype 2.2 Strain Bergen

Ulrik Fahnøe,[a] Louise Lohse,[a] Paul Becher,[b] Thomas Bruun Rasmussen[a]

DTU National Veterinary Institute, Technical University of Denmark, Lindholm, Kalvehave, Denmark[a]; EU and OIE Reference Laboratory for Classical Swine Fever (EURL), Institute of Virology, Department of Infectious Diseases, University of Veterinary Medicine, Hannover, Germany[b]

The complete genome sequence of the genotype 2.2 classical swine fever virus strain Bergen has been determined; this strain was originally isolated from persistently infected domestic pigs in the Netherlands and is characterized to be of low virulence.

Received 1 May 2014 **Accepted** 13 May 2014 **Published** 29 May 2014

Citation Fahnøe U, Lohse L, Becher P, Rasmussen TB. 2014. Complete genome sequence of classical swine fever virus genotype 2.2 strain Bergen. Genome Announc. 2(3): e00483-14. doi:10.1128/genomeA.00483-14.

Copyright © 2014 Fahnøe et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported license.

Address correspondence to Thomas Bruun Rasmussen, tbrur@vet.dtu.dk.

Classical swine fever virus (CSFV) belongs to the genus *Pestivirus* within the family *Flaviviridae*. CSFV is an important animal pathogen that causes disease in pig species and has low-, moderate-, and high-virulence characteristics (1). The CSFV genome consists of a positive-sense RNA, approximately 12.3 kb in length, which encodes a single polyprotein that is co- and posttranslationally cleaved to form the mature structural and nonstructural proteins. The CSFV strains can be divided into genotypes 1, 2, and 3, each comprising three to four subgenotypes (2, 3). CSFV strain Bergen represents a low-virulence strain that originally was isolated from persistently infected pigs in the Netherlands (4). The low-virulence phenotype of the Bergen strain was confirmed in a recent pathogenicity study (5). The Bergen strain has been grouped together with genotype 2.2 strains based on partial 5′-untranslated region (UTR) and E2 sequences (2, 3). Complete genomic sequences have been described for most CSFV genotypes. However, complete genome sequences from genotype 2.2 are lacking in the public sequence databases.

Here, we describe the complete genome sequence of the CSFV genotype 2.2 strain Bergen (isolate CSF0906) obtained from the CSFV collection at the EU Reference Laboratory (EURL). Viral RNA was extracted from infected PK15 cells, and full-length viral cDNAs were amplified by long reverse transcription-PCR (RT-PCR), as previously described (6), using cDNA primer 5′-GGGCCGTTAGGAAATTACCTTAGT- 3′ and PCR primers 5′-TCTATATGCGGCCGCTAATACGACTCACTATAGTATACGAGGTTAGTTCATTCTCGTGTACAATATTGGACAAACTAAATTCAGATTTGG-3′ and 5′-ATATGCGGCCGCGGGCCGTTAGGAAATTACCTTAGTCCAACTAT-3′. The sequencing library was generated from the RT-PCR product using the Ion Plus fragment library kit and sequenced using an Ion Torrent PGM (Life Technologies). Newbler (Roche) was used for *de novo* assembly and the Burrows-Wheeler Aligner (BWA) (7) for mapping of the reads using the *de novo* assembly as the reference sequence. Finally, consensus sequences were aligned using MAFFT in the Geneious software platform (Biomatters).

The final 12,295-nucleotide (nt)-long consensus sequence for the CSFV strain Bergen genome was obtained from a *de novo* assembly consisting of 16,318 sequence reads, with an average sequence depth of 268 reads per nt. The polyprotein-coding sequence is 11,697 nt long and contains 3,899 codons. The 5′ and 3′ UTRs are 372 and 226 nt long, respectively. A comparison with the previously published partial sequence (accession no. JQ411587, 3,508 nt) (3) comprising a part of the 5′ UTR and the coding sequence for the N-terminal autoprotease Npro, capsid protein C, envelope glycoproteins Erns, E1, E2, and the N-terminal part of p7 revealed 7 nt differences. These differences were all mapped to quasispecies populations in the deep sequencing data (data not shown). A comparison with a partial NS5B coding sequence (accession no. AF182909, 409 nt) revealed 100% identity. The complete genome sequence of CSFV strain Bergen should allow for further studies on the genetic diversity and the relationship between the CSFV genotype 2.1 and 2.2 strains.

Nucleotide sequence accession number. The genomic sequence of CSFV strain Bergen has been deposited in GenBank under the accession no. KJ619377.

ACKNOWLEDGMENT

This study was supported by DTU National Veterinary Institute.

REFERENCES

1. **Floegel-Niesmann G, Blome S, Gerss-Dülmer H, Bunzenthal C, Moennig V.** 2009. Virulence of classical swine fever virus isolates from Europe and other areas during 1996 until 2007. Vet. Microbiol. **139:**165–169. http://dx.doi.org/10.1016/j.vetmic.2009.05.008.
2. **Paton DJ, McGoldrick A, Greiser-Wilke I, Parchariyanon S, Song JY, Liou PP, Stadejek T, Lowings JP, Björklund H, Belák S.** 2000. Genetic typing of classical swine fever virus. Vet. Microbiol. **73:**137–157. http://dx.doi.org/10.1016/S0378-1135(00)00141-3.
3. **Postel A, Schmeiser S, Bernau J, Meindl-Boehmer A, Pridotkas G, Dirbakova Z, Mojzis M, Becher P.** 2012. Improved strategy for phylogenetic analysis of classical swine fever virus based on full-length E2 encoding sequences. Vet. Res. **43:**50. http://dx.doi.org/10.1186/1297-9716-43-50.
4. **Van Oirschot JT, Terpstra C.** 1977. A congenital persistent swine fever

infection. I. Clinical and virological observations. Vet. Microbiol. **2:**121–132. http://dx.doi.org/10.1016/0378-1135(77)90003-7.
5. **Lohse L, Nielsen J, Uttenthal A.** 2012. Early pathogenesis of classical swine fever virus (CSFV) strains in Danish pigs. Vet. Microbiol. **159:**327–336. http://dx.doi.org/10.1016/j.vetmic.2012.04.026.
6. **Rasmussen TB, Reimann I, Uttenthal A, Leifer I, Depner K, Schirrmeier H, Beer M.** 2010. Generation of recombinant pestiviruses using a full-genome amplification strategy. Vet. Microbiol. **142:**13–17. http://dx.doi.org/10.1016/j.vetmic.2009.09.037.
7. **Li H.** 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2. http://arxiv.org/abs/1303.3997v2.

Manuscript 3

Rescue of the highly virulent classical swine fever virus strain "Koslov" from cloned cDNA and first insights into genome variations relevant for virulence

Virology 468-470 (2014) 379–387

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Rescue of the highly virulent classical swine fever virus strain "Koslov" from cloned cDNA and first insights into genome variations relevant for virulence

Ulrik Fahnøe [a,b], Anders Gorm Pedersen [b], Peter Christian Risager [a], Jens Nielsen [a,1], Graham J. Belsham [a], Dirk Höper [c], Martin Beer [c], Thomas Bruun Rasmussen [a,*]

[a] DTU National Veterinary Institute, Technical University of Denmark, Lindholm, DK-4771 Kalvehave, Denmark
[b] Center for Biological Sequence Analysis, DTU Systems Biology, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
[c] Institute of Diagnostic Virology, Friedrich-Loeffler-Institut, Greifswald-Insel Riems, Germany

ARTICLE INFO

Article history:
Received 31 May 2014
Returned to author for revisions
7 July 2014
Accepted 21 August 2014

Keywords:
Flaviviridae
Pestivirus
Classical swine fever virus
CSFV
Koslov
Full-length cDNA clone
Virulence
Pathogenesis
Viral populations
Next generation sequencing

ABSTRACT

Classical swine fever virus (CSFV) strain "Koslov" is highly virulent with a mortality rate of up to 100% in pigs. In this study, we modified non-functional cDNAs generated from the blood of Koslov virus infected pigs by site-directed mutagenesis, removing non-synonymous mutations step-by-step, thereby producing genomes encoding the consensus amino acid sequence. Viruses rescued from the construct corresponding to the inferred parental form were highly virulent, when tested in pigs, with infected animals displaying pronounced clinical symptoms leading to high mortality. The reconstruction therefore gave rise to a functional cDNA corresponding to the highly virulent Koslov strain of CSFV. It could be demonstrated that two single amino acid changes (S763L and P968H) in the surface structural protein E2 resulted in attenuation in the porcine infection system while another single amino acid change within the nonstructural protein NS3 (D2183G) reduced virus growth within cells *in vitro*.

© 2014 Elsevier Inc. All rights reserved.

Introduction

Certain RNA viruses constitute important pathogens. These viruses can evolve rapidly and their genome replication is highly error-prone leading to significant diversity in virus properties including virulence. For example, the animal pathogen classical swine fever virus (CSFV), causing a highly contagious disease of domestic pigs and wild boar, can occur in a broad spectrum of variants that differ considerably in their properties; individual strains can display high, moderate or low virulence (Floegel-Niesmann et al., 2009). CSFV is a member of the *Pestivirus* genus within the family *Flaviviridae* and is closely related to the *Hepacivirus* genus, which contains the important human pathogen Hepatitis C virus and the novel equine and murine Hepaciviruses (Drexler et al., 2013). The CSFV strain "Koslov" is considered as one of the most virulent strains known and is an international standard for challenge experiments, with a mortality rate close to 100% (Bartak and Greiser-Wilke, 2000, Blome et al., 2012, Gabriel et al., 2012, Kaden and Lange, 2001, Kaden et al., 2001, Mittelholzer et al., 2000). Thus, a cDNA clone corresponding to the Koslov strain should be a valuable starting point for detailed analysis of the molecular determinants of virulence and other aspects of virus function.

In this study, we have generated full-length CSFV cDNA clones starting from the blood of a Koslov virus infected pig using a strategy that enables the production of numerous full-length cDNA clones directly from the viral RNA (Rasmussen et al., 2010, Rasmussen et al., 2013). However, RNA derived from each cDNA clone obtained initially was non-functional in terms of infectivity when tested in porcine cells. Based on sequence analysis of these non-functional cDNAs we were able to perform a reconstruction using site-directed mutagenesis so that RNA transcripts, encoding the consensus amino acid sequence were generated. These transcripts produced infectious virus when introduced into cells. Infection experiments in pigs proved that only the virus rescued

* Corresponding author. Tel./fax: +45 3588 7850.
E-mail address: tbrur@vet.dtu.dk (T.B. Rasmussen).
[1] Present address: Department of Microbiological Diagnostics & Virology, Statens Serum Institut, DK-2300 Copenhagen, Denmark.

http://dx.doi.org/10.1016/j.virol.2014.08.021
0042-6822/© 2014 Elsevier Inc. All rights reserved.

Table 1
Complete summary of all mutations found in the cloned cDNA used for reconstruction.

		KosA		KosB		KosC		KosD	
	Nucleotide position	Sequence	Effect	Sequence	Effect	Sequence	Effect	Sequence	Effect
N^{pro}	446	T		T		T		C	Y25H
	741	G		G		A	G123D	G	
	762	A		G	H130R	A		A	
	812	G		G		T	G147C	G	
C	996	T	T208M	C		C		C	
	1063	G	-	A		A		A	
	1068	G		A	G232D	G		G	
E^{rns}	1356	G		A	S328N	G		G	
	1494	T		A	V374D	T		T	
	1511	A		A		A		G	T380A
	1522	T	-	C		C		C	
	1843	T		T		T		C	-
E1	2122	A		A		G	-	A	
	2134	C		C		T	-	C	
	2191	A		G	-	A		A	
	2372	A		G	I667V	A		A	
	2386	A		G	-	A		A	
	2408	A		A		G	I679V	A	
E2	2617	T		C	-	T		T	
	2622	T		T		C	V750A	T	
	2661	T	S763L	T	S763L	C		C	
	2848	C		C		C		T	-
	2860	A		A		G	-	A	
	3205	A	-	A	-	G		G	
	3242					-	Frameshift		
	3262	C		T	-	C		C	
	3290	T		T		T		C	S973P
	3380	T		T		C	F1003L	T	
	3499	G		G		G		A	-
	3561	T		C	L1063S	T		T	
	3599	T		T		T		C	-
P7	3629	G		A	V1086I	G		G	
	3727	T	-	C		C		C	
	3818	A		A		A		G	I1149V
NS2-3	4032	A		A		G	Y1220C	A	
	4150	C	-	T		T		T	
	4255	T		T		C	-	T	
	4269	T		C	L1299P	T		T	
	4292	A		G	T1307A	A		A	
	4465	A	-	G		G		G	
	4606	T		T		T		C	-
	4612	C		G	-	G	-	C	
	4750	T		C	-	C	-	C	-
	4753	C		T	-	C		C	
	4987	A	-	T		T		T	
	5101	A	-	G		G		G	
	5166	T	T1598I	C		C		C	
	5170	A		G	-	A		A	
	5319			*CA*	Frameshift				
	5543	C	-	T		T		T	
	5691	A		G	N1773S	A		A	
	5771	A		A		A		G	K1800E
	6569	G	I2066V	A		A		A	

	6686	T		T		T		C	Y2105H
	6823	A	-	G		G		G	
	6921	A		A		G	D2183G	A	
NS4B	7837	G	-	A		A		A	
	8056	C	-	T		T		T	
	8213	T	P2614S	C		C		C	
	8487	A		A		A		T	E2705V
NS5A	8432	C	Y2687H	T		T		T	
	8510	G		A	A2713T	G		G	
	9745	T		C	-	T		T	
NS5B	9940	T		T		C	-	C	-
	10129	C		C		T	-	C	
	10289	T		T		T		C	C3306R
	10430	A		G	I3353V	A		A	
	10669	A		A		G	-	G	-
	11062	T		C	-	T		T	
	11065	A	M3564I	G		G		G	
	11071	A		G	-	A		A	
	11137	T		T		T		C	-
	11374	C	-	C	-	G		G	
	11599	C		C		C		T	-
	11623	A		A		G	-	A	
	11752	G	-	A		A		A	
	11983	A		A		A		G	-
3' UTR	12074	T	-	C		C		C	

At the left nucleotide position in the genome and the viral proteins are indicated. For each clone the sequenced nucleotide is shown and if in bold it differs from the consensus sequence. Additionally, the consequence for the amino acid sequence is shown where (-) is a silent mutation.

from the cDNA clone identical to the consensus sequence was as virulent as the parental Koslov strain. Attenuation was observed due to two single amino acid substitutions from the consensus sequence in the coding region for the glycoprotein E2.

Results and discussion

Reconstruction of Koslov cDNA clones

Genome length RT-PCR products (*ca.* 12.3 kb) were amplified using RNA isolated from the blood of a Koslov strain CSFV-infected pig on post inoculation day (PID) 6 and the amplicons were inserted into bacterial artificial chromosomes (BACs). The parental cDNA amplicon and independent cloned cDNAs were transcribed *in vitro* into RNA and tested for infectivity in PK15 cells using electroporation. Transcripts from the parental, uncloned, cDNA were found to give rise to infectious virus, whereas RNA from each of the tested cloned cDNAs obtained did not. Therefore, the complete CSFV cDNA cloned into four independent BACs (KosA, B, C and D) was sequenced. Each of these cDNA clones was found to contain a number of non-synonymous mutations and in two of the cDNAs frameshift mutations within the ORF were also detected (Table 1). We then proceeded to reconstruct the sequence of the parental strain, *i.e.*, the viral sequence corresponding, at the predicted amino acid level, to the Koslov strain consensus sequence (GenBank HM237795). To generate a cDNA encoding the parental amino acid sequence several rounds of site-directed mutagenesis were performed. The cDNA obtained after each round of site-directed mutagenesis was confirmed by full-length sequencing. In total, four reconstructed cDNAs (Kos_4aa, Kos_3aa, Kos_2aa and Kos) were produced (Fig. 1).

In vitro characterization of rescued viruses

The infectivity of the RNA transcripts obtained from the reconstructed cDNAs was investigated following their introduction into PK15 cells (Fig. 1). Each of the 4 different full-length RNAs was found to be infectious. Virus infectivity was determined by detection of the non-structural protein NS3, which was clearly seen in the cytoplasm of electroporated PK15 cells. In addition, the viruses rescued were capable of infecting other PK15 cells leading to the generation of high virus titers (Fig. 1).

The growth rates of the viruses rescued from each of the four cDNA clones, as well as from the uncloned Koslov amplicon, were further investigated. RNA from the three cDNAs that most resembled the consensus sequence (Kos_3aa, Kos_2aa and Kos) produced viruses with growth rates similar to the parental virus (Fig. 2*a*). The virus rescued from the fourth clone (Kos_4aa), however, displayed a much lower growth rate and had an additional amino acid substitution within the NS3-protein (D2183G). Subsequent analysis of RNA accumulation within infected cells by RT-qPCR did not demonstrate significant differences between the levels of viral RNA generated from each of these cDNAs (Fig. 2*b*). It could be that the mutation in the NS3 coding region, which encodes a key component in pestivirus replication (Lamp et al., 2013), is not affecting the replication rate of the viral genome directly, but is instead impairing assembly or release of the viruses (Moulin et al., 2007), leading to a drop in infectivity.

Testing of virulence in pigs

Viruses rescued from the intermediate Kos_3aa cDNA (termed vKos_3aa) and from the fully reconstructed Kos cDNA (termed vKos) were analyzed for their virulence in pigs. For each virus, three inoculated pigs were housed with two uninfected contact

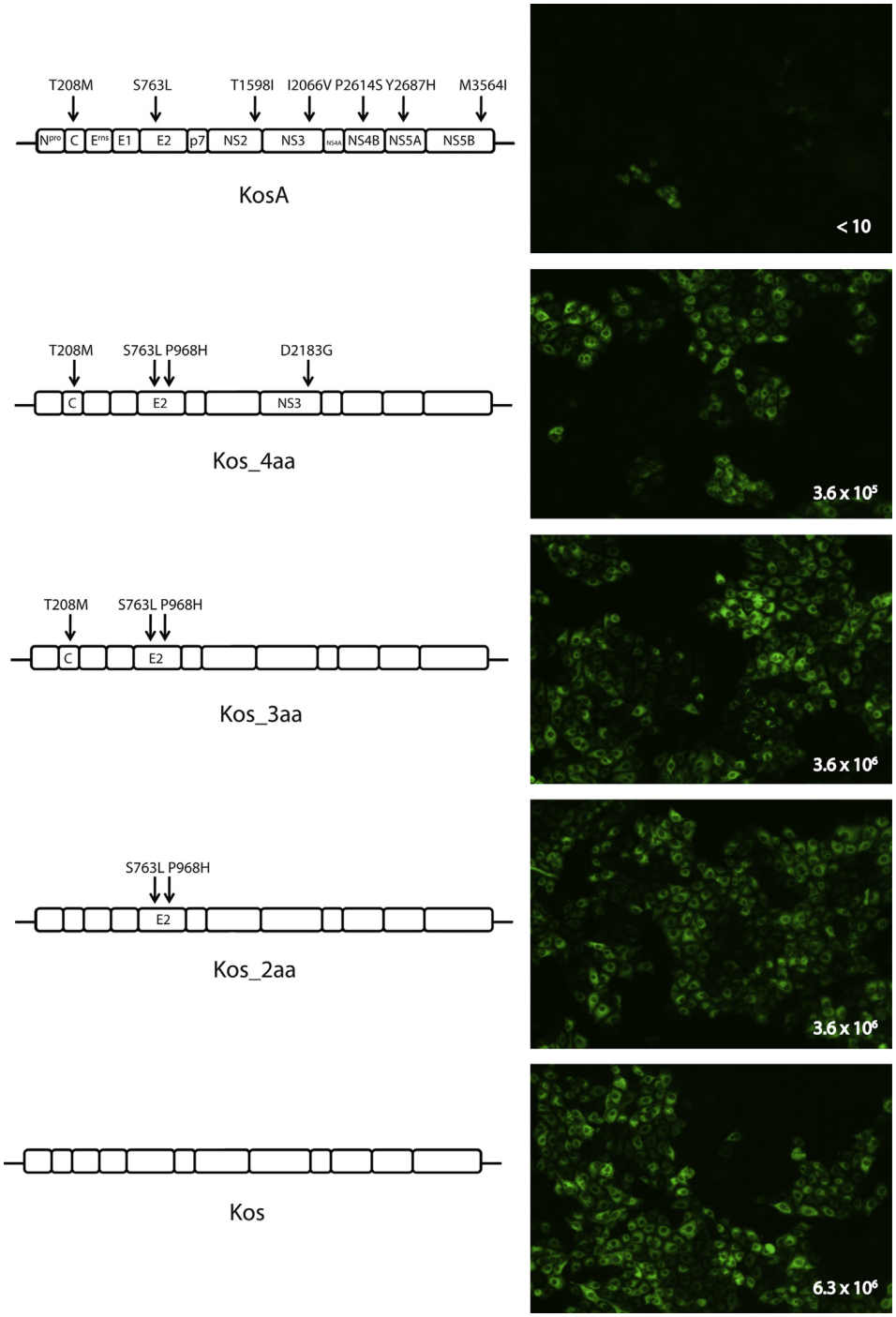


Fig. 1. Reconstruction of cDNAs and *in vitro* replication analysis. The figure shows the reconstruction by site-directed mutagenesis of the different CSFV Koslov cDNAs labeled on the left-hand side. Differences in predicted amino acid sequences between the constructs and the consensus sequence are indicated by arrows. Each new construct was tested for infectivity *in vitro* by introduction of RNA transcripts into PK15 cells and the production of the non-structural protein NS3 was visualized after 48 h using immunofluorescence staining of the cells. Rescued viruses were passaged once and the virus titers were determined and shown in the bottom right corner of each picture as $TCID_{50}$/ml.

pigs. Only the rescued virus corresponding to the consensus genome (vKos) produced a severe progression of disease in pigs and due to pronounced symptoms of classical swine fever (CSF), all three inoculated pigs were euthanized at PID5 (Fig. 3). Specifically, high fever (above 41 °C) was observed as early as PID3 in the inoculated pigs (Fig. 4*a*). This coincided with high viral RNA loads in both blood and nasal swabs (Fig. 4*b* and *c*) together with a drop in circulating B-cells (Fig. 4*d*). Platelet counts also fell during the infection (Fig. 4*e*). There was almost no delay in the onset of fever in the contact animals and this was also reflected in the changes in the levels of circulating B-cells (Fig. 4*d*). Already at PID6 both contact pigs had severe symptoms of CSF and had to be euthanized due to welfare reasons.

The intermediate virus (vKos_3aa) produced a delayed progression of disease. Two of the pigs inoculated with vKos_3aa were euthanized at PID10 due to continuing clinical disease (Fig. 3). The pigs developed high fever within the first week (Fig. 4*a*) and all infected pigs showed high viral RNA loads in blood and in nasal

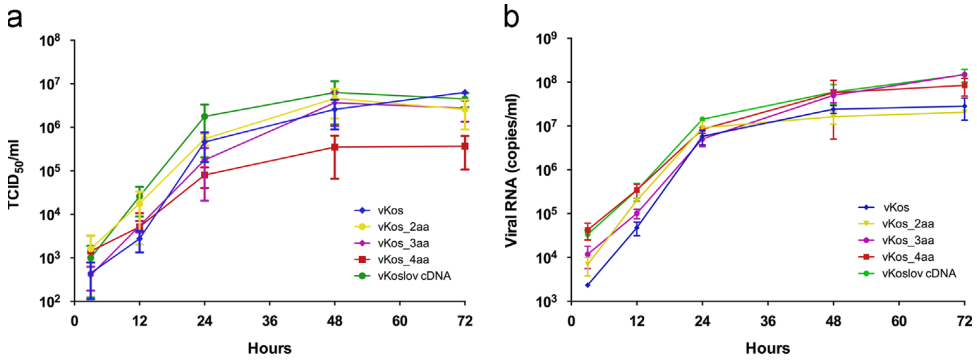


Fig. 2. Growth characteristics of viruses rescued from reconstructed cDNAs. Growth of the viruses in PK15 cells was measured by a) virus titration ($TCID_{50}$/ml) and b) RT-qPCR (viral RNA copies/ml) at 3, 12, 24, 48, and 72 h after infection. Means ± s.d. are shown for biological replicates ($n=3$).

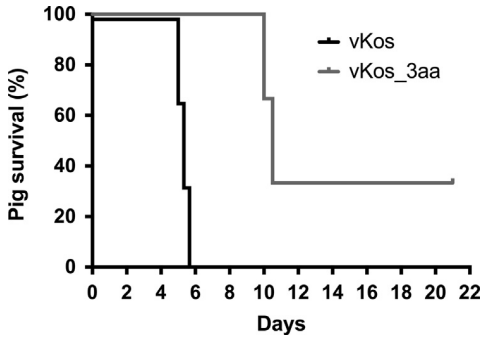


Fig. 3. Survival curves of pigs infected with viruses rescued from reconstructed cDNAs. Two groups of 3 pigs were inoculated with either vKos or vKos_3aa. The two groups showed significant differences in survival, Log rank (Mantel–Cox) test $p=0.0246$ and Gehan–Breslow–Wilcoxon test $p=0.0339$ ($n=3$).

swabs (Fig. 4b and Fig. 4c). A large drop in circulating B-cells and reductions in the platelet counts preceded the fever (Figs. 4d and e). One inoculated pig regained normal body temperature and increased platelet counts (Figs. 4a and e) but at necropsy (PID21) the presence of severe hemorrhages in the spleen and other internal organs indicated that the pig had not fully recovered from disease. The contact pigs developed fever and clinical signs around PID12, again preceded by reduction in B-cell counts (Fig. 4d). One contact pig was euthanized at PID18, because of severe persistent clinical signs. The fever of the other contact pig declined but the platelet counts dropped and the B-cell levels never recovered.

Replication *in vivo* was found to be more efficient for vKos compared to vKos_3aa, when measured as the level of viral RNA in blood (Fig. 4b). The RT-qPCR assays revealed virus in nasal swabs at PID4 in the vKos inoculated pigs, whereas this was observed one day later for the vKos_3aa inoculated animals (Fig. 4c).

From these experiments we conclude that vKos_3aa is moderately virulent, while vKos is significantly more virulent (Fig. 3), and has properties that closely resemble the highly virulent parental Koslov strain (Blome et al., 2012, Gabriel et al., 2012).

Molecular evolution of the viral populations

To monitor the molecular evolution of the viral populations *in vitro* and during infection of pigs, viral RNA from the inoculated viruses, as well as from the serum of each pig were deep sequenced. The deep sequencing revealed that all four inoculums were 100% identical to their respective cDNA sequence at the consensus level (data not shown), indicating that no adaptive mutations were needed for the rescued viruses to be functional. For vKos, we found no nt changes (compared to the cDNA sequence) at the consensus level in either the inoculum or the inoculated pigs (Table 2). However, this was not the case for vKos_3aa where we found that a methionine (M) codon, at position 208 in the structural C protein, completely reverted to the Koslov consensus sequence (encoding Threonine, T) after 10 days in all of the inoculated pigs, suggesting that this residue is important for viral fitness in the pig. However, this mutation alone, cannot explain the difference in virulence between the two tested reconstructed viruses, because it would then be expected that the contact pigs would exhibit more severe symptoms than the inoculated ones, and this was not apparent. It therefore appears that one, or both, of the other two amino acid substitutions (S763L and P968H) that are present in the surface structural protein E2, which is known to be required for attachment of the virus to cellular receptors and subsequent cell entry (El Omari et al., 2013), are also contributing to the difference in virulence between the viruses (vKos and vKos_3aa).

As mentioned above, the viruses in the vKos inoculated animals displayed no differences at the consensus sequence level. However, this was not the case for the contact animals in this group, where virus from one pig had 5 nucleotide substitutions in almost 100% of the population, two of which were missense mutations (Table 2). One substitution, L764P, is situated in the E2 protein, in a putative epitope (Chang et al., 2012), but the function is not known. Additionally, P764 is the predominant allele across all the full-length CSFV genomes found in Genbank (data not shown), but the leucine variant is present in several of the highly virulent genotype 1.1 strains; this indicates both residues are functional. The second missense mutation (G9821A) produced the substitution A3150T, this residue is positioned in the C-terminus of NS5A. This change did not seem to affect the virulence of the virus since this contact animal had as severe symptoms as the other pigs in that group (Fig. 6a–d).

A similar pattern could be observed for the contact pigs of the vKos_3aa group, which had three additional consensus changes by day 17–18, one of which was a missense mutation leading to the substitution E3390D (Table 2). This substitution is positioned at a conserved residue in the NS5B protein but did not appear to change the virulence of the population either.

SNP analyses were performed to further investigate the virus populations during infection. The inoculated viruses (vKos and vKos_3aa) displayed low-level variation (below 10%) scattered along the genome seen as synonymous and missense mutations. For the inoculated pigs the population distribution within the virus reflected the inoculum but this was not observed for the contact pigs. This is probably due to the transmission to the contacts being a bottleneck and subsequently a founder effect leading to a reduced amount of variation and genetic drift in the viral population compared to the inoculated pigs. The general tendency within the inoculated pigs in the vKos_3aa group was a

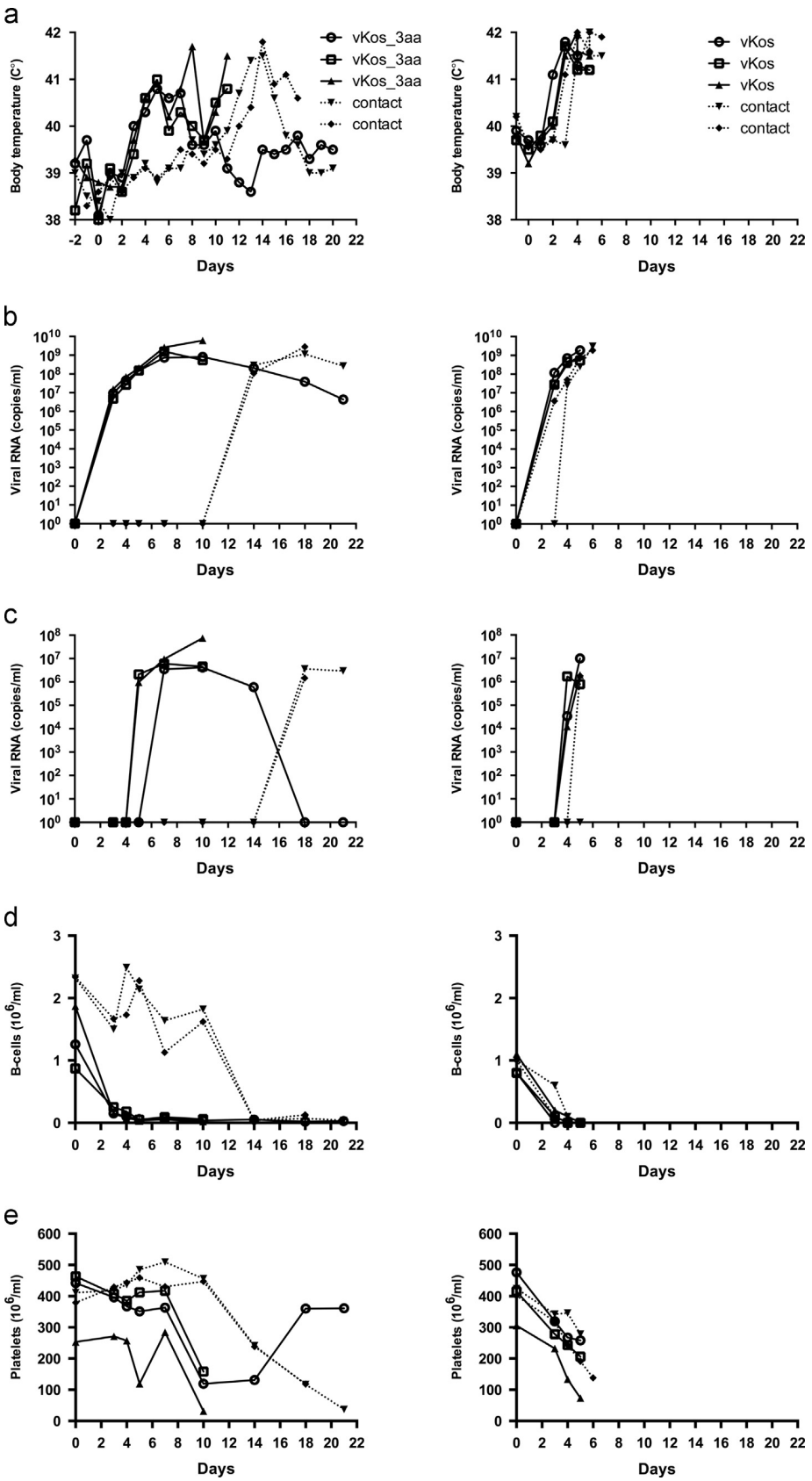


Fig. 4. Infection of pigs by reconstructed viruses. a) Body temperatures during the infection. b) Level of viral RNA in the blood measured by RT-qPCR. c) Level of viral RNA in nasal swabs measured by RT-qPCR. d) B-cell counts measured in EDTA blood and e) Platelet counts measured in EDTA blood.

Table 2
Summarized next generation sequencing results from animal experiments.

Virus	Type	PID	G279A Silent	T784C Silent	C996T T208M	C2452T Silent	C2661T S763L	T2664C L764P	C3276A P968H	A3307G Silent	G9821A A3150T	A10045G Silent	G10543T E3390D	T11680C Silent
vKos	Inoculum	0	-	-	-	-	-	-	-	-	-	-	-	-
			-	-	-	-	-	- (2%)	-	-	-	-	-	-
	Inoculated[a]	5	-	-	-	-	-	- (2%)	-	-	-	-	-	-
			-	-	-	-	-	-	-	-	-	-	-	-
	Contact[b]	6	-	-	-	-	-	-	-	-	-	-	-	-
			+	+	-	-	-	+	-	-	+	+	-	-
vKos_3aa	Inoculum	0	-	-	+	-	+	-	+	-	-	-	-	10%
			-	-	-	-	+	-	+	-	-	-	-	+
	Inoculated[a]	10	-	-	-	-	+	-	+	-	-	-	14%	+
			-	-	-	-	+	-	+	-	-	-	-	+
	Contact[b]	17	-	-	-	+	+	-	+	+	-	-	-	+
		18	-	-	-	-	+	-	+	-	-	-	+	+

Each (-) means that the variant cannot be detected in the sample and (+) indicates that the variant was detected in 100% in the viral population. The average sequence depth was above 1000 for all the samples.

[a] Representing serum samples from 3 inoculated pigs and () represents SNPs found in 1 nasal and one tonsil sample.
[b] Representing serum samples from 2 contact pigs.

Table 3
Primers used in this study.

Name	Sequence 5′–3′	Reference
CSF-Kos_Not1-T7-1-59	TCT ATA TGC GGC CGC TAA TAC GAC TCA CTA TAG TAT ACG AGG TTA GTT CAT TCT CGT ATG CAT GAT TGG ACA AAT CAA AAT TTC AAT TTG G	Modified from Leifer et al., 2010
CSF-kos_12313aR-NotI	ATA TGC GGC CGC GGG CCG TTA GGA AAT TAC CTT AGT CCA ACT GT	Modified from Leifer et al., 2010
CSF-kos_12313aR	GGG CCG TTA GGA AAT TAC CTT AGT CCA ACT GT	Leifer et al., 2010
3′CSF-kos_rev-RT	GGG CCG TTA GGA AAT TAC CTT AGT	Leifer et al., 2010
CSF-kos_5119F	ACA CCT TGG CTG GGT CCT TAG AGG	This study
CSF-Kos-6745-F	GCA CAG AGG TAC GGT ATT GAA GAT G	This study
CSF-Kos-7123-R	CTT GGT TTC CAG GGT CTG GCC AGT C	This study
CSF-Kos-962-F	AGA AGG ACA GCA GAA CTA AGC	This study
CSF-Kos-1547-R	TGG CCT GTT TCT GGC CTG GGT GAC C	This study
CSFV-Kos-1900-F	GTA CAC TAA CAA CTG CAC CCC GGC	This study
CSFV-Kos-3559-R	ACC AGC GGC GAG TTG TTC TGT TAG	This study

reduced amount of variation in the viral RNA extracted from serum compared to that from the inoculum and a lower frequency of each SNP was maintained in the population for both inoculated groups. A reason for this loss of variation in the pigs could be the selection pressure of the host compared to cell-culture leading to only some SNPs being retained. There was also a difference in the frequency and level of SNPs between the serum sample and the tonsil sample prepared from each pig post mortem, the tonsil samples showed more SNPs at higher frequencies than the serum samples (data not shown). The tonsils are known as the primary replication tissue for the CSFV virus and could maintain the variation of the different genomes, whereas serum will contain a mix of viruses derived from multiple tissues and not cell associated RNA genomes as in the tonsils.

As for the contact pig from the vKos group carrying the L764P substitution, it was possible to observe this mutation in the RNA at just above detection level in two of the inoculated pigs (Table 2). This could indicate a path of transmission, leading to fixation either by drift or by selection but clearly the pigs did not survive very long which could have enabled the virus population to change more significantly. Future studies, involving the modified Kos clone, will determine whether the L764P change does cause a change in the virulence and replication of this virus.

Interestingly, the silent mutation T11680C, which was found at a frequency of 100% in all pigs inoculated with vKos_3aa, was already present at 10% in the virus inoculum (Table 2). Since this mutation is silent it is likely to have been passively selected as a haplotype together with the C996T reversion. The reversion C996T is below the significant detection level of 1% in the inoculum (Radford et al., 2012), but a few reads in the sequence data did have this mutation indicating the presence of this variant (data not shown). Additionally, the missense mutation E3390D found in one of the contact pigs in the vKos_3aa group was also found in the serum of one inoculated pig at a frequency of 14% (Table 2), which could be the path of transmission for this virus. The two silent mutations found within the contact pig could not be detected in any of inoculated pigs; this may indicate a bottleneck effect of the transmission event leading to fixation of all three mutations.

Conclusions

In this work, we have successfully produced a fully functional CSFV cDNA corresponding to the highly virulent CSFV strain Koslov. The cDNA clone (Kos) will be the basis for future studies into virulence and viral population adaption. Indeed, as part of the reconstruction process, we have already identified several amino acid changes that significantly alter virus functionality and lower its virulence in the natural host animal. In addition, we have shown how this new clone can contribute towards understanding viral population adaptation and how it can be used to study the virulence of this important pathogen.

Materials and methods

Cells and viruses

The porcine kidney cell line PK15 (obtained from the Cell Culture Collection at the Friedrich-Loeffler-Institut, Germany) was propagated in cell culture medium containing 5% FCS. Blood

from a pig infected with CSFV strain "Koslov" (CSFV/1.1/dp/CSF0382/XXXX/Koslov) was obtained from Sandra Blome, Friedrich-Loeffler-Institut, Germany.

Generation of full-length cDNA

BACs containing full-length cDNAs corresponding to the genome of Koslov were obtained as previously described for the CSFV strains "Paderborn" and "C-strain Riems" (Rasmussen et al., 2010, Rasmussen et al., 2013). Briefly, viral RNA was purified from blood and full-length cDNA was amplified by RT-PCR using cDNA primer 3′CSF-kos_rev-RT and PCR primers CSF-Kos_Not1-T7-1-59 and CSF-kos_12313aR-NotI (Table 3) and subsequently inserted into *Not*I-digested pBeloBAC11 (New England Biolabs). The PCR primers were modified from those described previously (Leifer et al., 2010) by addition of *Not*I sites. Four independent BACs were named KosA, B, C and D (Table 1).

Reconstruction of cDNA

The BAC, Kos_4aa (GenBank KF977610), was obtained by site-directed mutagenesis using a megaprimer approach (Risager et al., 2013). Briefly, KosC was used as template for the megaprimer PCR with primers CSF-kos_5119F and CSF-kos_12313aR (Table 3), and the megaprimer, after gel purification, was used for the site-directed mutagenesis with KosA as vector backbone and downstream cloning in *E. coli* DH10B (Invitrogen, Carlsbad, USA).

Kos_3aa (GenBank KF977609) was obtained by the same protocol, using primers CSF-Kos-6745-F and CSF-Kos-7123-R and KosB as template for the megaprimer and Kos_4aa as vector backbone for the site-directed mutagenesis.

Subsequently, Kos_2aa (GenBank KF977608) was obtained using primers CSF-Kos-962-F and CSF-Kos-1547-R with KosC as template, and Kos_3aa as the vector backbone for the megaprimer PCR.

Finally, a fully reconstructed consensus clone (in terms of the encoded amino acid sequence) was produced using primers CSFV-Kos-1900-F and CSFV-Kos-3559-R. None of the clones initially isolated (as described above) had the Koslov consensus sequence in this stretch of the genome, therefore a full-genome cDNA product was used as template for the amplification of the megaprimer. Additionally, Kos_2aa was used as the vector backbone, which resulted in Kos (GenBank KF977607) being the only cDNA clone out of 15 with no predicted amino acid changes compared to the GenBank sequence (GenBank HM237795).

Rescue of virus and virus growth curves

BAC DNAs were purified from 4 ml overnight cultures of *E. coli* DH10B using a ZR BAC DNA Miniprep Kit (Zymo Research, Irvine, USA). BAC DNA (100-fold diluted) was used as template for full genome PCR amplification using primers CSF-Kos_Not1-T7-1-59 and CSF-kos_12313aR (Table 3). PCR products were purified with the Fermentas PCR purification kit and transcribed *in vitro* using a Megascript T7 kit (Invitrogen). Virus was rescued from RNA transcripts (1–5 μg) by electroporation of PK15 cells essentially as described previously (Friis et al., 2012). For visualization of virus-infected cells, NS3-specific murine antibodies (WB103/105, AHVLA Scientific, Surrey, UK) and Alexa 488 conjugated goat anti mouse IgG antibody (Molecular Probes, Carlsbad, USA) were used for immunofluorescence.

Virus growth curves were generated as previously described (Friis et al., 2012). Briefly, PK15 cells were infected at an MOI of 0.1 $TCID_{50}$/cell and incubated for up to three days. At 3, 12, 24, 48 and 72 h post infection, cell samples were harvested for virus titration and RT-qPCR.

Full-genome sequencing

PCR products amplified from viral cDNA or BACs were consensus sequenced using a Genome Sequencer FLX (Roche, Mannheim, Germany) or an Ion PGM (Life Technologies, Carlsbad, USA). Both Newbler (Genome Sequencer Software suite; Roche) and BWA (Li, 2013) were used for mapping the reads to the consensus sequence of CSFV strain Koslov (GenBank Accession number HM237795). Application of a combination of Samtools (Li et al., 2009) and Lo-Freq-snp-caller (Wilm et al., 2012) was used for downstream single nucleotide polymorphism (SNP) analysis together with SnpEff (Cingolani et al., 2012). Finally, clone consensus sequences were aligned using MAFFT in Geneious (Biomatters, Auckland, New Zealand).

Animal infections

Crossbred pigs (8–10 weeks old) obtained from the high health status swine herd at DTU were used for the animal infections. Three pigs (for each virus) were inoculated with either vKos or vKos_3aa via the intranasal route with a defined dose (10^6 $TCID_{50}$/pig) and two in-contact pigs in the same pen were mock-inoculated with cell culture medium. The inocula were back-titrated to confirm the inoculation dose. Body temperature and clinical signs were monitored daily. At pre-defined days (PID 0, 3, 4, 5, 7, 10, 14, 18 and 21) EDTA-blood and serum samples were collected for virological, hematological, and immunological examination as previously described (Nielsen et al., 2010). Furthermore, nasal swabs were obtained on the same sampling days and the viral RNA from these and from EDTA-blood was purified using a MagNA Pure LC total nucleic acid isolation kit (Roche). The level of viral RNA was determined by RT-qPCR as described previously (Rasmussen et al., 2007).

Experimental procedures and animal management protocols were carried out in accordance with the requirements of the Danish Animal Experimentation Inspectorate.

References

Bartak, P., Greiser-Wilke, I., 2000. Genetic typing of classical swine fever virus isolates from the territory of the Czech Republic. Vet. Microbiol. 77, 59–70.
Blome, S., Aebischer, A., Lange, E., Hofmann, M., Leifer, I., Loeffen, W., Koenen, F., Beer, M., 2012. Comparative evaluation of live marker vaccine candidates "CP7_E2alf" and "flc11" along with C-strain "Riems" after oral vaccination. Vet. Microbiol. 158, 42–59.
Chang, C.Y., Huang, C.C., Deng, M.C., Huang, Y.L., Lin, Y.J., Liu, H.M., Lin, Y.L., Wang, F. I., 2012. Antigenic mimicking with cysteine-based cyclized peptides reveals a previously unknown antigenic determinant on E2 glycoprotein of classical swine fever virus. Virus Res. 163, 190–196.
Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6, 80–92.
Drexler, J.F., Corman, V.M., Muller, M.A., Lukashev, A.N., Gmyl, A., Coutard, B., Adam, A., Ritz, D., Leijten, L.M., van Riel, D., Kallies, R., Klose, S.M., Gloza-Rausch, F., Binger, T., Annan, A., Adu-Sarkodie, Y., Oppong, S., Bourgarel, M., Rupp, D., Hoffmann, B., Schlegel, M., Kummerer, B.M., Kruger, D.H., Schmidt-Chanasit, J., Setien, A.A., Cottontail, V.M., Hemachudha, T., Wacharapluesadee, S., Osterrieder, K., Bartenschlager, R., Matthee, S., Beer, M., Kuiken, T., Reusken, C., Leroy, E. M., Ulrich, R.G., Drosten, C., 2013. Evidence for novel hepaciviruses in rodents. PLoS Pathog. 9, e1003438.
El Omari, K., Iourin, O., Harlos, K., Grimes, J.M., Stuart, D.I., 2013. Structure of a pestivirus envelope glycoprotein E2 clarifies its role in cell entry. Cell. Rep. 3, 30–35.
Floegel-Niesmann, G., Blome, S., Gerss-Dulmer, H., Bunzenthal, C., Moennig, V., 2009. Virulence of classical swine fever virus isolates from Europe and other areas during 1996 until 2007. Vet. Microbiol. 139, 165–169.
Friis, M.B., Rasmussen, T.B., Belsham, G.J., 2012. Modulation of translation initiation efficiency in classical swine fever virus. J. Virol. 86, 8681–8692.
Gabriel, C., Blome, S., Urniza, A., Juanola, S., Koenen, F., Beer, M., 2012. Towards licensing of CP7_E2alf as marker vaccine against classical swine fever-Duration of immunity. Vaccine 30, 2928–2936.

Kaden, V., Lange, B., 2001. Oral immunisation against classical swine fever (CSF): onset and duration of immunity. Vet. Microbiol. 82, 301–310.
Kaden, V., Schurig, U., Steyer, H., 2001. Oral immunization of pigs against classical swine fever. Course of the disease and virus transmission after simultaneous vaccination and infection. Acta Virol. 45, 23–29.
Lamp, B., Riedel, C., Wentz, E., Tortorici, M.A., Rumenapf, T., 2013. Autocatalytic cleavage within classical swine fever virus NS3 leads to a functional separation of protease and helicase. J. Virol. 87, 11872–11883.
Leifer, I., Hoffmann, B., Hoper, D., Bruun Rasmussen, T., Blome, S., Strebelow, G., Horeth-Bontgen, D., Staubach, C., Beer, M., 2010. Molecular epidemiology of current classical swine fever virus isolates of wild boar in Germany. J. Gen. Virol. 91, 2687–2697.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. 1000 genome project data processing subgroup. Bioinformatics 25, 2078–2079.
Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM 2013 arXiv:1303.3997v2[q-bio.GN] Please note that this reference is a preprint hosted at arXiv.org.
Mittelholzer, C., Moser, C., Tratschin, J.D., Hofmann, M.A., 2000. Analysis of classical swine fever virus replication kinetics allows differentiation of highly virulent from avirulent strains. Vet. Microbiol. 74, 293–308.
Moulin, H.R., Seuberlich, T., Bauhofer, O., Bennett, L.C., Tratschin, J.D., Hofmann, M. A., Ruggli, N., 2007. Nonstructural proteins NS2-3 and NS4A of classical swine fever virus: essential features for infectious particle formation. Virology 365, 376–389.
Nielsen, J., Lohse, L., Rasmussen, T.B., Uttenthal, A., 2010. Classical swine fever in 6- and 11-week-old pigs: hematological and immunological parameters are modulated in pigs with mild clinical disease. Vet. Immunol. Immunopathol. 138, 159–173.
Radford, A.D., Chapman, D., Dixon, L., Chantrey, J., Darby, A.C., Hall, N., 2012. Application of next-generation sequencing technologies in virology. J. Gen. Virol. 93, 1853–1868.
Rasmussen, T.B., Risager, P.C., Fahnøe, U., Friis, M.B., Belsham, G.J., Höper, D., Reimann, I., Beer, M., 2013. Efficient generation of recombinant RNA viruses using targeted recombination-mediated mutagenesis of bacterial artificial chromosomes containing full-length cDNA. BMC Genomics 14, 819.
Rasmussen, T.B., Reimann, I., Uttenthal, A., Leifer, I., Depner, K., Schirrmeier, H., Beer, M., 2010. Generation of recombinant pestiviruses using a full-genome amplification strategy. Vet. Microbiol. 142, 13–17.
Rasmussen, T.B., Uttenthal, A., Reimann, I., Nielsen, J., Depner, K., Beer, M., 2007. Virulence, immunogenicity and vaccine properties of a novel chimeric pestivirus. J. Gen. Virol. 88, 481–486.
Risager, P.C., Fahnoe, U., Gullberg, M., Rasmussen, T.B., Belsham, G.J., 2013. Analysis of classical swine fever virus RNA replication determinants using replicons. J. Gen. Virol. 94, 1739–1748.
Wilm, A., Aw, P.P., Bertrand, D., Yeo, G.H., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., Nagarajan, N., 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40, 11189–11201.

Manuscript 4

Analyzing the fitness of a viral population: Only a minority of circulating virus haplotypes are viable

Analyzing the fitness of a viral population: Only a minority of circulating virus haplotypes are viable

Ulrik Fahnøe[a,b], Anders Gorm Pedersen[b], Carolin Dräger[c], Richard J Orton[d,e], Sandra Blome[c], Dirk Höper[c], Martin Beer[c], Thomas Bruun Rasmussen[a,#]

[a]*DTU National Veterinary Institute, Technical University of Denmark, Lindholm, DK-4771 Kalvehave, Denmark*

[b]*Center for Biological Sequence Analysis, DTU Systems Biology, Technical University of Denmark, Denmark*

[c]*Institute of Diagnostic Virology, Friedrich-Loeffler-Institut, Greifswald-Insel Riems, Germany*

[d]*Institute of Biodiversity, Animal Health, and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, UK*

[e]*MRC – University of Glasgow Centre for Virus Research, Institute of Infection, Inflammation and Immunity, College of Medical, Veterinary and Life Sciences, University of Glasgow, UK*

#Corresponding author:

Mailing address: DTU National Veterinary Institute, Technical University of Denmark, Lindholm, DK-4771 Kalvehave, Denmark

Phone: +45 3588 7850. Fax: +45 3588 7850. E-mail: tbrur@vet.dtu.dk

Abstract

Positive strand RNA viruses are among the most common pathogens affecting both animals and humans. These viruses can evolve at a high rate due to their error-prone replication machinery, resulting in a cloud of closely related haplotypes with potentially different properties. Here we present the first extensive experimental and computational analysis of a series of full-length cDNA clones derived from a single population of classical swine fever virus. Using a reverse genetics approach, we were able to study the fitness and functionality of each cDNA clone. Only 12% of the full-length cDNA clones yielded RNA transcripts that were infectious *in vitro*. Sequencing of the cloned cDNAs revealed an association between the number of missense mutations (compared to their consensus) and functionality. Deep sequencing of the parental cDNA population was in agreement with the SNP distribution seen in the cDNA clones and allowed assessment of how well a number of haplotype reconstruction algorithms performed.

Thus, the majority of circulating virus haplotypes are non-functional. However, since these virus particles must by necessity be descendants of functional ancestors, we hypothesized that it should be possible to produce an infectious form of the virus by reconstructing ancestral sequences corresponding to selected internal nodes in the cDNA clone phylogeny. Specifically, we reconstructed the two major inferred haplotypes of the population using site-directed mutagenesis. Both haplotypes proved infectious in cell culture, but displayed distinct phenotypes *in vitro* and *in vivo*.

Introduction

RNA viruses are among the most common pathogens affecting both animals and humans. Classical swine fever virus (CSFV) is an example of a positive-strand RNA virus and is the causative agent of classical swine fever (CSF), which is economically important because of its highly contagious nature in pigs and wild boar (Vandeputte and Chappuis 1999). Different strains of the virus can vary greatly in the severity of symptoms within infected animals, with a virulence range classified as low, moderate or high (Floegel-Niesmann et al. 2009). The most recent outbreaks of CSF in Europe were mainly caused by genotype 2.3 viruses, which display moderate virulence (Kaden et al. 2004; Pol et al. 2008; Postel et al. 2013). One of the latest field isolates is the German CSFV strain "Roesrath", which has been found to be moderately virulent in domestic pigs and wild boar (Leifer et al. 2010; Petrov et al. 2014). The genotype 2.3 strain would be a suitable model for this important group of viruses, but so far no functional cDNA clone has been established from CSFV "Roesrath" or other genotype 2.3 strains.

CSFV and many other RNA viruses have high mutation rates compared to eukaryotic organisms due to the lack of proofreading by their RNA dependent RNA polymerase (RdRp)(Drake 1993). The high mutation rate is beneficial to the virus in the sense that it enables rapid adaptation and escape from host antiviral responses, but it is also potentially harmful because the mutations may be detrimental or even lethal for the virus. The fidelity of the RdRp has been shown to affect virulence with increased fidelity leading to attenuation and restricted diversity in the viral population (Zeng et al. 2014). Generally, a high mutation rate will cause a population to consist of a swarm of different, but closely related, haplotypes forming a flat fitness landscape that makes the population more robust to mutations (Lauring and Andino 2010). It has been proposed that viral populations with a high diversity of haplotypes will have a better chance of surviving host responses to infection (Bonhoeffer and Nowak 1997). For CSFV, high virulence has been related to high diversity (Töpfer et al. 2013).

The analysis of viral haplotype composition has typically relied on clonal approaches and subsequent sequencing of partial genomes. In recent years, however, Next-Generation Sequencing (NGS) has rapidly become the preferred technology to address this issue. Using NGS it is possible to sequence a large number of individuals in a population of viral genomes at potentially great depths (meaning that many different versions of a given genomic region

will be seen). This allows for detection of single nucleotide polymorphisms (SNP) present at low frequencies within a virus population. However, the linkage of SNPs in separate parts of the genome is not easily discerned from SNP analysis of NGS data because of read length limitations (unless in close proximity it is not known if a SNP in one read derives from the same viral genome as a SNP in another read). In order to address this question, several computational tools have been created, which use different approaches for reconstructing haplotypes and quasispecies from NGS data (Zagordi et al. 2011; Prosperi and Salemi 2012). Benchmarking of such methods has indicated relatively poor performance with full-length viral genomes (Prosperi et al. 2013; Schirmer et al. 2014) and it was concluded that additional datasets are needed in order to test these methods.

Attempts have been made to overcome sequencing errors introduced by NGS technology and use the data to model the fitness landscape of a virus population (Acevedo et al. 2014). However, all described NGS approaches have limitations concerning the assessment of the functionality of each individual haplotype.

Here we present a study of the evolution of a single viral population using a combination of deep sequencing and analysis of multiple, individual, full-length cloned cDNAs. Specifically, 70 cloned cDNAs representing individual haplotypes within the virus population were fully sequenced and their phenotypes (in terms of replication efficiency and infectivity) assessed in cell culture. From the full-length sequences, both haplotype compositions and phylogeny were determined. Haplotype prediction tools were applied to the NGS data and compared to the cDNA clones. Furthermore, ancestral reconstruction of sequences corresponding to certain internal nodes in the phylogeny was performed. Finally, viruses rescued from the reconstructed ancestral cDNAs were used for infection of pigs, in order to determine their virulence in comparison to the parental CSFV “Roesrath” isolate. This is, to our knowledge, the first study that combines deep sequencing and functional analysis of cloned cDNAs representing the spectrum of haplotypes to investigate evolution of a viral population.

Results

Eighty-four unique full-length cDNAs cloned into bacterial artificial chromosome (BAC) vectors were generated from RT-PCR products obtained using RNA extracted from a fifth passage of the CSFV "Roesrath" isolate (termed CSFV_Roesrath_P5). RNA transcripts were produced from individual cloned cDNAs and were tested for replication competence in PK-15 cells. Transcripts from 15 of the cDNA clones (18%) were scored as functional and replicated in PK-15 cells whereas 69 (82%) were non-functional without any indication of RNA replication (fig. 1A). For the 15 replicating cDNA clones differences in replication efficiency were observed with varying phenotypes ranging from all cells becoming infected to only a few small foci of infected cells. In order to address these differences, harvests from cells displaying CSFV protein production following introduction of the viral RNA transcripts were passaged once on PK-15 cells to establish whether infectious virus had been produced. Ten cDNA clones (12%) were identified as producing virus progeny after this additional passage and were classified as "infectious" whereas the cDNAs yielding non-infectious RNA transcripts, although producing detectable viral protein, were termed "replication competent" (fig. 1A).

The full-length consensus sequences of the cDNA clones (70 out of the 84 cDNA clones in total) were determined in order to identify haplotypes within the viral population. This included all of the infectious and replicating cDNA clones as well as the majority (55 out of 69) of the non-functional cDNA clones. The consensus sequences of each individual cDNA clone were determined from *de novo* assembly of the reads. All cDNA clones had unique sequences when compared to the CSFV "Roesrath" reference sequence (GU233734) and to each other, with an average of 11.4 mutations, compared to the reference. Out of all observed SNPs in the coding sequence, 38% were silent whilst 62% resulted in amino acid (aa) differences ("missense mutations"). Indels (1-2 nt) were only observed for some of the non-functional cDNAs (12 out of the 55 non-functional cDNA clones). All indels in the coding sequence resulted in frameshifts. Using the above phenotypic classification, the different types of mutations could be assigned to the three different classes (infectious, replication competent and non-functional) of the cDNA clones (fig. 1B). A significant difference in the total number of mutations could be observed between the infectious and non-functional variants. Thus, a significantly higher number of missense mutations were present in non-functional cDNA clones compared to those that yielded infectious RNAs, whereas the number of silent mutations was approximately the same in the two classes. The replication competent (but

non-infectious) transcripts had an intermediate level of missense mutations between the infectious and non-functional cDNA clones. This distribution of mutations indicates an association between the number of missense mutations and a decrease in functionality. Further analysis revealed that only the NS5B protein (RdRp) had significantly less missense mutations in the infectious compared to the non-functional cDNA clones ($p= 0.003$; data not shown). Apart from that, missense mutations were fairly evenly distributed.

Phylogenetic analysis, selection dN/dS analysis and NGS SNP analysis

The complete set of unique cDNA clones could be used to assess the molecular quasispecies structure together with the haplotype variants present in the viral population. First, the phylogenetic tree for the cDNA clones was inferred using Bayesian methods (fig. 2). The unrooted tree structure was mostly star-like, with a majority of cDNA sequences branching out from a single deep node. However, the tree did display some structure, with a few distinct monophyletic groups, one of which (marked in red) included a subgroup (marked in purple). Labeling individual leafs (representing individual cDNA clones) according to functionality, showed that sequences belonging to the infectious and replication competent variants were distributed over most of the subgroups and that functional variants were present in all of the major groups. The leaves for the infectious variants were found to be closer to the internal nodes than those that were non-functional, in agreement with the mutation distribution (fig. 1B). For three of the monophyletic subgroups (red, purple and blue) linked mutations were observed with two groups having five mutations and one having three. For example, the red subgroup has three linked mutations (two missense mutations in NS2 and NS4B and a silent mutation in NS5B).

In order to investigate selective pressure in the viral population, we estimated dN/dS ratios based on the cDNA clone sequences and the phylogeny (table 1). A number of different models (corresponding to different hypotheses about the presence of negative and positive selection) were fitted to the data. The model (Model 7) with the highest weight probability (w) was the one that best fitted the data. In addition, increasing the number of categories did not increase the probability. The best fitting model suggested that all sites had dN/dS rates <1 indicating that all sites in the sequences had experienced mostly neutral and negative selection.

The full-length cDNA, which was obtained from CSFV_Roesrath_P5 and used as input for the cloning procedure, was deep sequenced by NGS using the FLX platform to investigate the SNP distribution within the parental un-cloned viral population. Simultaneously, an independent full-length RT-PCR product obtained from the same cDNA was deep sequenced using the Ion PGM platform. Table 2 shows the SNP distribution from these data, after rigorous error correction and filtering. All observed SNPs displayed frequencies below 20% and the consensus sequences for each product were identical to the CSFV "Roesrath" reference (GU233734). Almost all those mutations that were present in more than one of the cloned cDNAs, could also be detected in the SNP data. Two of the groups with one SNP that was not detected in the NGS data could be found in an alignment bam file but was not deemed to be significant by either Lofreq or V-phaser 2 (data not shown). Unique mutations in the cDNA clones might be caused by last round replication errors, or they may be low frequency SNPs in the populations (<1%). Subsequently, the focus was put on mutations appearing in more than one cDNA clone. Color-coding of individual SNPs (table 2) indicates how they are linked on the different haplotypes and is comparable to the branch colors of the phylogeny (fig. 2). By comparing this data to the phylogeny, 3 groups of haplotypes with more than one SNP could be recognized in the tree. These three groups are colored blue, red and purple (the latter being a subclade of the red group) in fig. 2. The purple and the blue haplotype groups had similar frequencies (about 5%) within the population. However, by looking at SNP frequencies alone, it would not be possible to separate these two groups, and only in combination with the phylogeny of the cDNA clones could these haplotypes be distinguished (fig. 3).

Haplotype and quasispecies reconstruction from NGS data and benchmarking of tools

Several prediction tools have been developed for reconstructing quasispecies and haplotypes of viral populations based on NGS data. The unique cDNA sequence dataset generated in this study can be used to assess how well each prediction tool performs since we have both NGS data and full-length sequences from individual cDNA clones reflecting the existing haplotypes. The haplotype prediction tools QuRe, ShoRah and PredictHaplo were applied to the FLX data set and gave rise to 10, 89 and 4 haplotypes respectively. In order to determine how well each tool performed, the predictions were aligned to the cDNA clone

sequences and the alignments were used to build phylogenies by Bayesian inference. Each consensus tree could then be evaluated with respect to how well the predictions fitted into the monophyletic groups of the cDNA clone population (fig. 4A-C) and thereby determine whether the haplotypes were correctly predicted. The assumption is here that the set of 70 cDNA clones sequenced give a representative picture of the main groupings in the entire virus population. QuRe gave the best result by predicting 3 of the clades correctly and having a precision of 30% (fig 4B). This tool was able to predict both the deepest node and the red group with approximately the same population proportions indicated by the cDNA clones. ShoRah and PredictHaplo managed to predict only the deepest node in the phylogeny with a precision of 25% and 1.1% respectively. In particular, ShoRah seemed to predict a lot of false haplotypes all of which were far from internal nodes (fig. 4A), while PredictHaplo underestimated the number of haplotypes by only predicting four haplotypes close to the deepest node (fig. 4C). QuRe, ShoRah and PredictHaplo identified 21%, 7% and 7% of the monophyletic groups respectively, under the assumption that the 14 clades identified within the cDNA clones from the phylogeny covered all major haplotypes. None of the predictors were able to reconstruct either the blue or the purple haplotypes. In addition, even though QuRe recognized the red group, it was not able to reconstruct the full genome and the predictions were truncated at both the 5′ and 3′ untranslated region (UTR) termini (data not shown).

Ancestral reconstruction of internal nodes

As mentioned above, the majority of cloned viral sequences were non-functional. However, the closer a cDNA clone was to an internal node in the phylogeny the more likely it was to be functional, an observation also made by other groups (Pybus et al. 2007). In particular the number of missense mutations separating a sequence from the deepest ancestor seemed to be important for determining functionality (fig. 1B). Since each circulating virus must be the descendant of a functional ancestor, we suggest that ancestral reconstruction of sequences corresponding to internal nodes will lead to fully functional and infectious virus variants. We therefore performed ancestral reconstruction using PAML (Yang 1997), which employs a maximum likelihood method to perform computational reconstruction of sequences corresponding to internal nodes in the phylogeny (fig. 5A). The

ancestral predictions are shown for the four major groups, and contain both silent and missense mutations compared to the consensus sequence.

Testing of reconstructed cDNA clones *in vitro*

We decided to produce constructs corresponding to the inferred ancestral sequences at the black and red diamond shaped nodes in figure 5A to test their functionality. This was done using site-directed mutagenesis, removing mutations step by step, to produce a cDNA clone identical to the ancestral sequence at the black diamond shaped node (here termed "Ros"; fig. 5B). RNA transcripts derived from Ros proved infectious in PK-15 cells. Indeed, growth curves showed that the virus rescued from Ros (termed "vRos") proliferates at least as well as the virus rescued from the parental cDNA (vRos_cDNA) in cell culture (fig. 6). From the Ros cDNA clone the ancestral sequence at the red node was subsequently constructed using two steps of site-directed mutagenesis. This sequence, named "Ros_S1359N_A2668T", had the two missense mutations (S1359N in NS2 and A2668T in NS4B), but not the silent NS5B mutation (T11992C)(fig. 5B). Each step added one missense mutation and transcripts containing each of the individual changes (Ros_S1359N and Ros_A2668T) were infectious in PK-15 cells (data not shown). The final construct Ros_S1359N_A2668T also proved to be infectious in cell culture and the rescued virus was termed vRos_S1359N_A2668T. As both ancestral reconstructions led to infectious viruses, we decided to test their replication efficiency in cell culture. PK-15 cells were infected with the same infectious dose and RNA was extracted at 2, 8 and 12 hours and the level of CSFV genomes then measured by RT-qPCR. This analysis showed that the vRos_S1359N_A2668T replicated significantly faster than vRos (fig. 7). This could be observed at both 8 and 12 hours post infection.

Deep sequencing of rescued viruses in comparison to the parental virus

Viruses (vRos and vRos_S1359N_A2668T) rescued from the two reconstructed ancestral cDNAs were deep sequenced and compared to CSFV_Roesrath_P5 (table 2) used for the initial cloning in order to identify mutations and adaptations to cell culture. Furthermore, a 2nd passage of the same virus isolate (CSFV_Roesrath_P2) was deep sequenced. Fig. 8 depicts the

SNP distribution across the cDNAs mapped to the CSFV "Roesrath" reference sequence (GU233734). In the vRos and CSFV_Roesrath_P2 sequences only low frequency SNPs could be observed in each population and both had consensus sequences identical to the deepest ancestral node except for a silent SNP (T7792C) at 64% for the CSFV_Roesrath_P2 sample. The vRos_S1359N_A2668T had a very similar pattern of SNPs as the vRos but with the two missense mutations G4449A and G8375A fixed at 100% as expected. The mutations leading to the aa substitutions S1359N and A2668T were observed in the population of CSFV_Roesrath_P5 at about 18% (table 2) but not even as low frequency SNPs in CSFV_Roesrath_P2.

Testing of viruses rescued from reconstructed cDNAs *in vivo*

To further investigate the observed differences in replication rate *in vitro* between the two reconstructed viruses (fig. 7) these viruses were also tested in the natural host. The two rescued viruses were analysed in parallel with the CSFV_Roesrath_P2 to compare their virulence. Three groups of weaner pigs, each with five animals, were inoculated with the vRos, CSFV_Roesrath_P2 and vRos_S1359N_A2668T respectively. Mild clinical symptoms were observed in the groups infected with vRos and CSFV_Roesrath_P2 (fig. 9A) and there were growth impediments for a few of the pigs in each group. Two pigs were euthanized at day 12 post infection from the group inoculated with vRos due to persistent clinical symptoms and declining general health. No disease symptoms were observed for the group inoculated with vRos_S1359N_A2668T. It should be noted that the CSFV strain "Roesrath" is moderately virulent and has been shown to cause highly variable clinical pictures, from subclinical (Mouchantat et al. 2014), to acute-lethal and chronic (Petrov et al. 2014). All three groups seroconverted against CSFV by day 28 of the experiment, and anti-CSFV antibodies could be detected as early as day 14. All pigs were scored as positive in a specific neutralisation test after day 14 (data not shown). Viremia was monitored using RT-qPCR assays to measure viral RNA in blood samples taken during the course of the experiment. The groups inoculated with vRos and CSFV_Roesrath_P2 had an almost identical profile of viremia except on day 14 where a higher load of vRos was apparent in the bloodstream (fig. 9B). For both groups viremia declined after day 14, but was detectable at low levels until the end of the experiment on day 42. However, the group infected with vRos_S1359N_A2668T had much lower levels of viral

RNA in the blood, viremia peaking between day 7 and 10, and virus being undetectable after day 14. Indeed, lower levels of viremia were observed for the group infected with vRos_S1359N_A2668T compared to groups inoculated with vRos and CSFV_Roesrath_P2 at every sample time examined. Oral swabs were taken on the same days as blood samples and assayed for viral RNA in the oral cavity using RT-qPCR (fig. 9C). Only the groups infected with CSFV_Roesrath_P2 and vRos had measurable viral RNA loads in the swab samples after day 21, and continued being positive at low levels until the end of the experiment. Taken together, vRos and CSFV_Roesrath_P2 showed similar characteristics *in vivo* whereas the vRos_S1359N_A2668T displayed an attenuated phenotype.

Discussion

In this study we present a large collection of unique full-length cDNA clones derived from a single RNA virus population. Seventy of these cDNA clones were fully sequenced and this unique data set allowed us to perform detailed investigations of the quasispecies and haplotype distributions. We also determined phenotypes of RNA transcripts derived from every single cDNA clone. Finally, we reconstructed ancestral viruses corresponding to internal nodes in the virus phylogeny, which resulted in virus progeny with different phenotypes both *in vitro* and *in vivo* in the natural host system.

Previous studies of RNA virus population structure with respect to population diversity and haplotype distribution have been based on pure NGS approaches or on partial genomic sequences. To the best of our knowledge this is the first study that combines NGS haplotyping with full-length cDNA clone data. Importantly, each cDNA clone was also investigated for its functionality in cell culture. Surprisingly, we found that 88% of the investigated cDNA clones were non-infectious. The number of missense mutations of each cDNA clone (compared to the consensus sequence) was found to be correlated with functionality.

The phylogeny derived from the cloned cDNA sequences revealed a diverse population including several monophyletic groups containing missense mutations, which gave rise to a "star-like" appearance of the inferred phylogenetic tree. Use of dN/dS analysis revealed that the population was under mild negative selection pressure. Deep sequencing of the parental un-cloned RT-PCR product confirmed the presence in the population, of the majority of mutations, which were also seen in more than one cDNA clone. Despite the fact that some of the mutations in each clone may be due to errors by the RT-PCR, comparison of the NGS data from the two RT-PCRs showed similar SNP frequencies, which indicated that these clones could be used to study the population structure. The phylogenetic structure showed several groups with linked mutations, so-called haplotypes. Haplotypes found in the cDNA clones were also detected in the SNP analysis for the deep sequenced virus population (CSFV_Roesrath_P5). However, the NGS data, in conjunction with haplotype prediction tools, was not sufficient to elucidate the finer haplotype structure of the population as detected by the reconstruction analysis. All three tools performed relatively poorly with QuRe as the best performing (since it identified the two major groups). Our data therefore confirms that

haplotype reconstruction from NGS data using these tools is still not very precise as has been reported by other groups (Prosperi et al. 2013; Schirmer et al. 2014). These methods have recently been shown to give good results but only for very diverse populations with many linked SNPs along each haplotype (Di Giallonardo et al. 2013; Giallonardo et al. 2014). Until the technology improves, haplotype reconstruction results from NGS data using such tools should be taken as indicative only and not as a true representation of the population.

As described above, the closer the cDNA clone sequence was to an internal node, the more likely it was to yield infectious RNA transcripts. Whenever a sample of viral RNA is cloned, then each cDNA will represent the last round of replication and only a small proportion, in our case 12%, may be infectious. However, each cDNA clone must be derived from RNAs originating from an infectious ancestor deeper in the phylogeny. We here used computational methods to infer the ancestral sequence corresponding to a number of internal nodes in the phylogeny. The two major ancestors, Ros and Ros_S1359N_A2668T, were reconstructed and tested in cell culture and in pigs. Both were found to be infectious in PK-15 cells thereby confirming our hypothesis that viruses corresponding to internal nodes must be functional. The vRos virus replicated at a lower rate than vRos_S1359N_A2668T in cell culture indicating that the S1359N (in NS2) and A2668T (in NS4B) substitutions are of importance for replication efficiency of the virus. These substitutions have not been described before and both positions are fully conserved in all sequenced CSFV isolates present in GenBank (data not shown). In order to test the virulence of each construct, both viruses were used together with CSFV_Roesrath_P2 in the natural host animal. The vRos was found to be as virulent as the original isolate CSFV_Roesrath_P2, whereas vRos_S1359N_A2668T displayed an attenuated phenotype in pigs. The difference between the *in vitro* and the *in vivo* replication efficiencies suggests that the mutations in vRos_S1359N_A2668T virus could be adaptations specific for cell culture. The S1359N and A2668T haplotype was not detected in deep sequencing of CSFV_Roesrath_P2, thereby confirming that these mutations were not part of the original isolate population diversity.

Concluding remarks

This study combines a variety of approaches to explore viral population structure in depth. NGS alone is limited to SNP calling and reconstruction of haplotypes, but the latter is

mostly useful in highly heterogeneous populations, where overlaps between individual sequencing reads are abundant. We have shown that the additional use of full genome sequencing of cDNA clones can provide powerful new data, which can be used to infer haplotype structure, explore phenotypes, and identify mutations of interest. We suggest that these approaches with great benefit could be applied to other RNA viruses, thereby leading to better understanding of the viral population dynamics of these important pathogens.

Materials and Methods

Virus isolates

The CSFV strain "Roesrath" was used for the experiments (CSFV/2.3/wb/CSF1045/2009/Roesrath; Genbank accession number GU233734). Two different cell culture passages derived from the same isolate were used: the CSFV_Roesrath_P5, which was a fifth passage sample from PK-15 cells whereas the CSFV_Roesrath_P2 was a second passage sample of the same isolate.

Generation of cloned cDNAs

The cloned cDNAs were produced from CSFV RNA as described previously (Fahnøe et al. 2014). In short, viral RNA was extracted from CSFV_Roesrath_P5 by a combined Trizol/RNeasy protocol. Subsequently, the viral genomes were amplified by RT-PCR to generate full-length genome amplicons flanked by *Not*I sites and with a T7 promoter upstream of the cDNA sequence using primers CSFV-Ros_Not1-T7-1-59 and CSFV-Ros_12313aR_Not1 (table 3). Fragment termini were digested with *Not*I and products were ligated into the bacterial artificial chromosome (pBeloBAC11). Individual colonies were propagated on LB plates supplemented with chloramphenicol, and screened following restriction enzyme digestion. PCR products for the *in vitro* transcription and the sequencing were obtained from each cloned cDNA using same forward primer and CSFV-Ros_12313aR (table 3).

Testing of RNA transcripts from full-length cDNA clones *in vitro*

cDNAs were transcribed using a Megascript T7 RNA transcription kit. Run-off RNA transcripts were electroporated into porcine PK-15 cells and incubated at 37°C in Eagles medium with 5% FCS. After 72 hours plates were stained for the presence of pestivirus antigens using biotinylated pig anti-CSFV/BVDV polyclonal IgG followed by avidin-conjugated horseradish peroxidase (eBioscience) for detection of viral proteins using microscopy. Cell supernatants from the replication competent transcripts were passaged onto uninfected PK-15 cells and incubated for a further 72 hours.

Determination of haplotypes from cloned cDNAs

Consensus sequences for the PCR products obtained from the cloned cDNAs were determined using the Ion PGM platform (Life technologies) or by using a Miseq instrument (Illumina). Additionally, the RT-PCR products obtained from the sample used for the cloning were deep sequenced with the FLX genome sequencer (Roche) and the Ion PGM platform. Sequencing data were assembled by the Newbler *de novo* assembler (Roche) and mapped to the CSFV "Roesrath" reference sequence (GU233734) by the BWA aligner using the BWASW algorithm (Li and Durbin 2010) and processed by Samtools (Li et al. 2009). Consensus sequences were aligned using the MAFFT algorithm in Geneious R7. The FLX and Ion PGM data was corrected for homopolymer errors by the RC454 tool using 454 settings (Henn et al. 2012). This tool integrates the Mosaic aligner (Lee et al. 2014) for mapping the reads to GU233734. Samtools were applied for bam file processing and SNPs were called by V-Phaser2 and Lofreq for comparison (Wilm et al. 2012; Yang et al. 2013). Subsequently, the SnpEffect tool was used to determine SNP effects (Cingolani et al. 2012).

Phylogenetic analysis, dN/dS analysis and ancestral reconstruction of internal nodes

cDNA clone sequences aligned by MAFFT were compared to the consensus sequence and mutations were categorised as silent, missense, deletions or situated within the 5′ UTR or 3′

UTR using Geneious R7. T-tests were performed in Graphad Prism 6.0.e. The alignment was run through jModelsTest 2.1.5 to determine the best substitution matrix and the General time reversible (GTR) outperformed the others. Phylogeny was constructed using MrBayes v3.2.1 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) on a full-length cDNA sequence alignment (GTR, nst=6). The Markov chain Monte Carlo algorithm was run for 20,000,000 iterations, with a sampling frequency of 14400, and using three independent chains in order to check for convergence. Burn in was set at 25% of samples. The consensus tree was visualized in FigTree v.1.4.0.

For dN/dS analysis, the CodeML program from the PAML package was used (Yang 1997; Yang 2007). In order to perform the dN/dS analysis, an alignment of the cDNA clone ORFs was generated by MAFFT in Geneious R7. These ORFs were then analysed using the CodeML module of PAML. A number of different models (corresponding to different hypotheses about the presence or absence of positive or negative selection) were fitted to the data. Specifically we investigated the models NSsites = 3, 7, 11, or 12 in combination with NcatG = 3, 5, or 9. The best performing model was determined by using AIC (Akaike Information Criterion), which were calculated based on the maximized likelihoods from each individual model.

Ancestral reconstruction of the internal nodes was performed using PAML. The BaseML program was applied on the full-length nucleotide alignment using GTR as substitution model. The internal node sequences were aligned by MAFFT in Geneious R7.

Haplotype and quasispecies reconstruction from NGS data and performance assessment

After homopolymer error correction by RC454, cleaned reads were processed by the following predictors: QuRe v9.9994 (Prosperi and Salemi 2012), ShoRAH v0.5.1 (Zagordi et al. 2011) and PredictHaplo v1.0 (Roth V http://bmda.cs.unibas.ch/HivHaploTyper/) for full-length haplotype reconstruction according to the developer's recommendations. NGS data from the FLX platform only were applied. Predicted haplotypes were aligned to the cDNA clone sequences by MAFTT in Geneious R7 separately. Again, the phylogeny was constructed using MrBayes v3.2.1 on a full-length cDNA-sequence alignment as described above.

Reconstruction of haplotypes by site-directed mutagenesis

The reconstruction of cDNA clones was performed as previously described (Fahnøe et al. 2014). Two rounds of site-directed mutagenesis using a megaprimer approach generated the deepest node in the phylogeny and the consensus cDNA clone Ros. Briefly, the megaprimer was generated using CSFV-Ros-1180-F and CSFV-Ros-2107-R as primers and Ros16B as template. The purified PCR product was used as megaprimer for megaPCR with Ros35C as template. Second round PCR was performed with CSFV-Ros-8093-F and CSFV-Ros-9435-R as primers and Ros16B as template for the megaprimer. Subsequently, Ros35C.2 was used as template for the megaPCR generating Ros35C.2.1, which was renamed Ros being a 100% match to the consensus sequence and the deepest node in the phylogeny. Two rounds of site-directed mutagenesis also generated Ros_S1359N_A2668T that corresponds to the node of the largest monophyletic subgroup. Initially, primer CSFV-Ros-4018-F and CSFV-Ros-4599-R were used together with Ros9C as template for the megaprimer. Ros was used as template together with the megaprimer for the MegaPCR that generated Ros_S1359N. The last round megaprimer was obtained with CSFV-Ros-8093-F and CSFV-Ros-9435-R as primers and Ros9C as template, which was used along with Ros_S1359N as template to produce Ros_S1359N_A2668T. This clone was shown by full-length sequencing to only include the two missense mutations (S1359N and A2668T) compared to Ros. Sequences for individual cDNA clones (e.g. Ros9C, Ros16B, Ros35C, Ros35C.2, Ros35C.2.1) can be obtained upon request.

Testing of virus rescued from reconstructed cDNAs *in vitro*

Virus growth was determined by growth curves as previously described (Friis et al. 2012). Briefly, rescued virus was used to infect PK-15 cells (MOI 0.1 TCID50/cell) and cultured for 3 days. At 3, 12, 24, 48 and 72 hours total RNA was isolated after a freeze thaw cycle of cells with medium and the CSFV genome copy numbers were measured by RT-qPCR (Hoffmann et al. 2005). The viral replication assay was adapted from (Tamura et al. 2012). In short, 300,000 PK-15 cells were infected at an MOI of 1.5 TCID50/cell and RNA from cells was harvested at 2, 8 and 12 hours, and the level of viral RNA was measured as described above.

Testing of rescued viruses *in vivo*

Fifteen weaner pigs were divided into 3 groups of 5 pigs that each received a different inoculum (Group 1, vRos; Group 2, CSFV_Roesrath_P2; Group 3, vRos_S1359N_A2668T). The pigs were inoculated using a vaporizer device (2 ml oral and nasal) with 10^5 $TCID_{50}$/100 µl. Body temperature and clinical signs were observed on a daily basis. Blood samples and oral swabs were taken at days 0, 2, 4, 7, 10, 14, 21, 28 and 42 post-infection. The IDEXX CSFV antibody ELISA (IDEXX, Bern, Switzerland) and neutralisation tests were performed on the serum samples as previously described (Gabriel et al. 2012). Viral RNA was detected using CSFV RT-qPCR protocols for both blood and oral swab samples as described (Hoffmann et al. 2005).

Acknowledgements

This work was supported by the European project Epi-SEQ (research project supported under the 2nd Joint Call for Transnational Research Projects by EMIDA ERA-NET [FP7 project no. 219235]) and by the German Federal Ministry for Education and Research (BMBF, grant 01KI1016A).

References

Acevedo A, Brodsky L, Andino R. 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505:686-690.

Bonhoeffer S, Nowak MA. 1997. Pre-existence and emergence of drug resistance in HIV-1 infection. Proc. Biol. Sci. 264:631-637.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80-92.

Di Giallonardo F, Zagordi O, Duport Y, Leemann C, Joos B, Kunzli-Gontarczyk M, Bruggmann R, Beerenwinkel N, Gunthard HF, Metzner KJ. 2013. Next-generation sequencing of HIV-1 RNA genomes: Determination of error rates and minimizing artificial recombination. PLoS One 8:e74249.

Drake JW. 1993. Rates of spontaneous mutation among RNA viruses. Proc. Natl. Acad. Sci. U. S. A. 90:4171-4175.

Fahnøe U, Pedersen AG, Risager PC, Nielsen J, Belsham GJ, Höper D, Beer M, Rasmussen TB. 2014. Rescue of the highly virulent classical swine fever virus strain "koslov" from cloned cDNA and first insights into genome variations relevant for virulence. Virology 468-470C:379-387.

Floegel-Niesmann G, Blome S, Gerss-Dulmer H, Bunzenthal C, Moennig V. 2009. Virulence of classical swine fever virus isolates from Europe and other areas during 1996 until 2007. Vet. Microbiol. 139:165-169.

Friis MB, Rasmussen TB, Belsham GJ. 2012. Modulation of translation initiation efficiency in classical swine fever virus. J. Virol. 86:8681-8692.

Gabriel C, Blome S, Urniza A, Juanola S, Koenen F, Beer M. 2012. Towards licensing of CP7_E2alf as marker vaccine against classical swine fever - duration of immunity. Vaccine 30:2928-2936.

Giallonardo FD, Töpfer A, Rey M, Prabhakaran S, Duport Y, Leemann C, Schmutz S, Campbell NK, Joos B, Lecca MR et al. 2014. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res. 42:e115.

Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf CM, Ryan EM, Gnerre S et al. 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 8:e1002529.

Hoffmann B, Beer M, Schelp C, Schirrmeier H, Depner K. 2005. Validation of a real-time RT-PCR assay for sensitive and specific detection of classical swine fever. J. Virol. Methods 130:36-44.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

Kaden V, Lange E, Polster U, Klopfleisch R, Teifke JP. 2004. Studies on the virulence of two field isolates of the classical swine fever virus genotype 2.3 rostock in wild boars of different age groups. J. Vet. Med. B Infect. Dis. Vet. Public Health 51:202-208.

Lauring AS, Andino R. 2010. Quasispecies theory and the behavior of RNA viruses. PLoS Pathog. 6:e1001005.

Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. 2014. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One 9:e90581.

Leifer I, Hoffmann B, Höper D, Rasmussen TB, Blome S, Strebelow G, Horeth-Bontgen D, Staubach C, Beer M. 2010. Molecular epidemiology of current classical swine fever virus isolates of wild boar in germany. J. Gen. Virol. 91:2687-2697.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics 26:589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079.

Mouchantat S, Globig A, Böhle W, Petrov A, Strebelow HG, Mettenleiter TC, Depner K. 2014. Novel rope-based sampling of classical swine fever shedding in a group of wild boar showing low contagiosity upon experimental infection with a classical swine fever field strain of genotype 2.3. Vet. Microbiol. 170:425-429.

Petrov A, Blohm U, Beer M, Pietschmann J, Blome S. 2014. Comparative analyses of host responses upon infection with moderately virulent classical swine fever virus in domestic pigs and wild boar. Virol. J. 11:134.

Pol F, Rossi S, Mesplede A, Kuntz-Simon G, Le Potier MF. 2008. Two outbreaks of classical swine fever in wild boar in france. Vet. Rec. 162:811-816.

Postel A, Moennig V, Becher P. 2013. Classical swine fever in Europe - the current situation. Berl. Munch. Tierarztl. Wochenschr. 126:468-475.

Prosperi MC, Salemi M. 2012. QuRe: Software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics 28:132-133.

Prosperi MC, Yin L, Nolan DJ, Lowe AD, Goodenow MM, Salemi M. 2013. Empirical validation of viral quasispecies assembly algorithms: State-of-the-art and challenges. Sci. Rep. 3:2837.

Pybus OG, Rambaut A, Belshaw R, Freckleton RP, Drummond AJ, Holmes EC. 2007. Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. Mol. Biol. Evol. 24:845-852.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Schirmer M, Sloan WT, Quince C. 2014. Benchmarking of viral haplotype reconstruction programmes: An overview of the capacities and limitations of currently available programmes. Brief Bioinform 15:431-442.

Tamura T, Sakoda Y, Yoshino F, Nomura T, Yamamoto N, Sato Y, Okamatsu M, Ruggli N, Kida H. 2012. Selection of classical swine fever virus with enhanced pathogenicity reveals synergistic virulence determinants in E2 and NS4B. J. Virol. 86:8602-8613.

Töpfer A, Höper D, Blome S, Beer M, Beerenwinkel N, Ruggli N, Leifer I. 2013. Sequencing approach to analyze the role of quasispecies for classical swine fever. Virology 438:14-19.

Vandeputte J, Chappuis G. 1999. Classical swine fever: The European experience and a guide for infected areas. Rev. Sci. Tech. 18:638-647.

Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40:11189-11201.

Yang X, Charlebois P, Macalalad A, Henn MR, Zody MC. 2013. V-phaser 2: Variant inference for viral populations. BMC Genomics 14:674-2164-14-674.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586-1591.

Yang Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555-556.

Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. 2011. ShoRAH: Estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics 12:119-2105-12-119.

Zeng J, Wang H, Xie X, Li C, Zhou G, Yang D, Yu L. 2014. Ribavirin-resistant variants of foot-and-mouth disease virus: The effect of restricted quasispecies diversity on viral virulence. J. Virol. 88:4008-4020.

Table 1. Test of selection models.

Model (NSsites)	Categories (NcatG)	K	lnL	AIC	ΔAIC	w
7	3	92	-21579.8570	43343.7139	0	0.3082
7	5	92	-21579.8590	43343.7180	0.0041	0.3076
7	9	92	-21579.8784	43343.7569	0.0430	0.3017
3	3	95	-21579.6190	43349.2379	5.5240	0.0195
12	9	95	-21579.8186	43349.6371	5.9232	0.0160
12	5	95	-21579.8340	43349.6680	5.9541	0.0157
12	3	95	-21579.8445	43349.6889	5.9750	0.0155
11	3	95	-21579.8451	43349.6901	5.9762	0.0155
3	5	99	-21579.6190	43357.2379	13.5240	3.5657E-4
3	9	107	-21579.6190	43373.2379	29.5240	1.1962E-7
11	5	95	-21650.0458	43490.0916	146.3777	5.0506E-33
11	9	95	-21749.7222	43689.4443	345.7304	2.5969E-76

K is the number of free parameters for each model, lnL is the maximized log likelihoods, AIC is the Akaike weight criterion and w is the model probability.

Table 2. SNP analysis of the viral population (CSFV_Roesrath_P5) by NGS deep sequencing

Protein	Nucleotide position	Roesrath reference (GU233734)	Variant	SNP Total (%)	SNP FLX (%)	SNP PGM (%)	Variant (aa)	Haplotype in cDNA clones
N^{pro}	787	T	C	1.1	0.9	1.2	Silent	Present
C	919	C	T	1.2	0.9	1.3	Silent	Not found
	938	C	T	4.2	4.2	4.2	Silent	Present
	1058	A	T	5.0	5.1	4.9	**T229S**	
E^{rns}	1673	G	A	4.1	4.5	3.9	**G434R**	Present
	1750	C	T	2.2	2.7	2	Silent	Not found
E2	2459	G	A	4.6	3.8	4.8	**D696N**	
	3559	T	G	4.2	3.5	4.4	Silent	
p7	3622	G	A	2.2	2.3	(2.2)	Silent	Not found
	3637	C	T	2.8	3.1	2.8	Silent	Present
NS2-3	3847	C	A	2.1	2.3	2	Silent	Present
	4171	A	G	1.8	(1.4)	1.9	Silent	Not found
	4449	G	A	18.2	18.3	18.2	**S1359N**	
	5593	C	G	3.5	(3.5)	(5.1)	Silent	
	5998	T	C	2.7	3.2	2.5	Silent	Not found
	6541	T	C	4.1	4.5	4	Silent	Present
	6592	T	A	5	4.2	5.3	Silent	
	6781	C	T	4.8	4.4	5	Silent	
NS4B	8220	A	G	2.4	(2.1)	2.5	**K2616R**	Not found
	8302	T	C	10.8	9.1	11.2	Silent	
	8375	G	A	18.2	18.1	18.2	**A2668T**	
NS5B	9958	T	C	4.8	4.4	4.9	Silent	
	10915	T	C	1.1	(0.7)	1.1	Silent	Present
	11101	A	G	3.4	3	3.5	Silent	Not found
	11407	C	T	3.0	3	3	Silent	Not found
	11533	A	G	2.4	2.3	2.4	Silent	Present
	11572	C	T	2.4	2.4	2.4	Silent	Present
	11575	C	T	3.8	4.2	3.7	Silent	Present
	11992	T	C	18.3	18.2	18.4	Silent	

Two independent PCR products obtained from the viral population (CSFV_Roesrath_P5) were NGS deep sequenced by the FLX and Ion PGM platforms. The SNP frequencies are shown, as percentages, for both datasets combined called by both Lofreq and V-Phaser 2 (SNP Total), followed by the FLX and Ion PGM SNP percentages respectively. SNPs frequencies called only with V-phaser 2 are shown in parenthesis. The last column depicts the SNPs association to the haplotypes in the cDNA clones in fig. 2, in which the colors in the column correspond to the colors in the phylogeny.

Table 3. Primers used in this study

Primer	Sequence 5′ – 3′
CSFV-Ros_Not1-T7-1-59	5′-TCA TAT GCG GCC GCT AAT ACG ACT CAC TAT AGT ATA CGA GGT TAG CTC GTT CTC GTA TAC GAT ATC GGA TAC ACT AAA TTT CG-3′
CSFV-Ros_12313aR_Not1	5′-ATA TGC GGC CGC GGG CCG TTA GGA AAT TAC CTT AGT CCA ACT GT-3′
CSFV-Ros_12313aR	5′-GGG CCG TTA GGA AAT TAC CTT AGT CCA ACT GT-3′
CSFV-Ros_cDNA	5′-GGGCCGTTAGGAAATTACCTTAGT-3′
CSFV-Ros-8093-F	5′-GGAGCTGTAGCAGCCCACAATGC-3′
CSFV-Ros-4018-F	5′-GACTTGGCTACAGTACCTCGTCAGC-3′
CSFV-Ros-4599-R	5′-GAAACGAGGTTGGTCCCACCAGC-3′
CSFV-Ros-1180-F	5′-CCAGCCCGTGGCAGCCGAGAAC-3′
CSFV-Ros-2107-R	5′-CAGGTTCTTCGTGGGACTGGGGG-3′
CSFV-Ros-8512-F	5′-CAATCAGCTGGGCCCCCGCC-3′
CSFV-Ros-9435-R	5′-CCCCAGTATCAGTACCGAGGGCC-3′

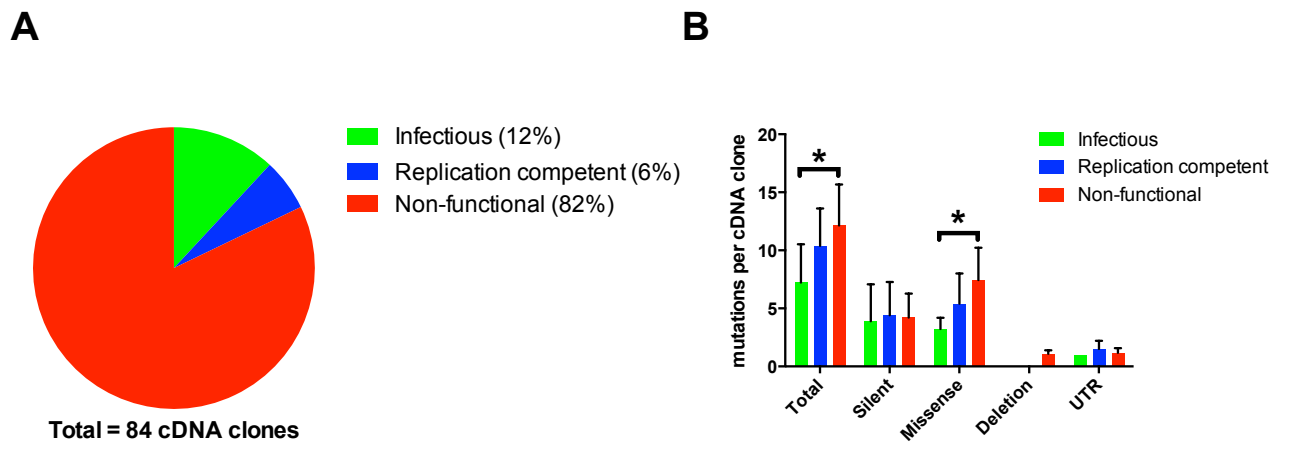


Fig. 1. Phenotypic and mutational distribution of the cDNA clones. A) Functionality of the cDNA clones according to the functional classes; "Infectious", "Replication competent" and "Non-functional". B) Distribution of mutation effects according to the different types of mutations; "Silent", "Missense", "Deletion" and "untranslated region (UTR)". Means ± s.d. are shown. The T-test was applied to determine significance difference between the infectious and non-funtional cDNA clones total (p= 0.0002) and missense (p= 3.3E-5).

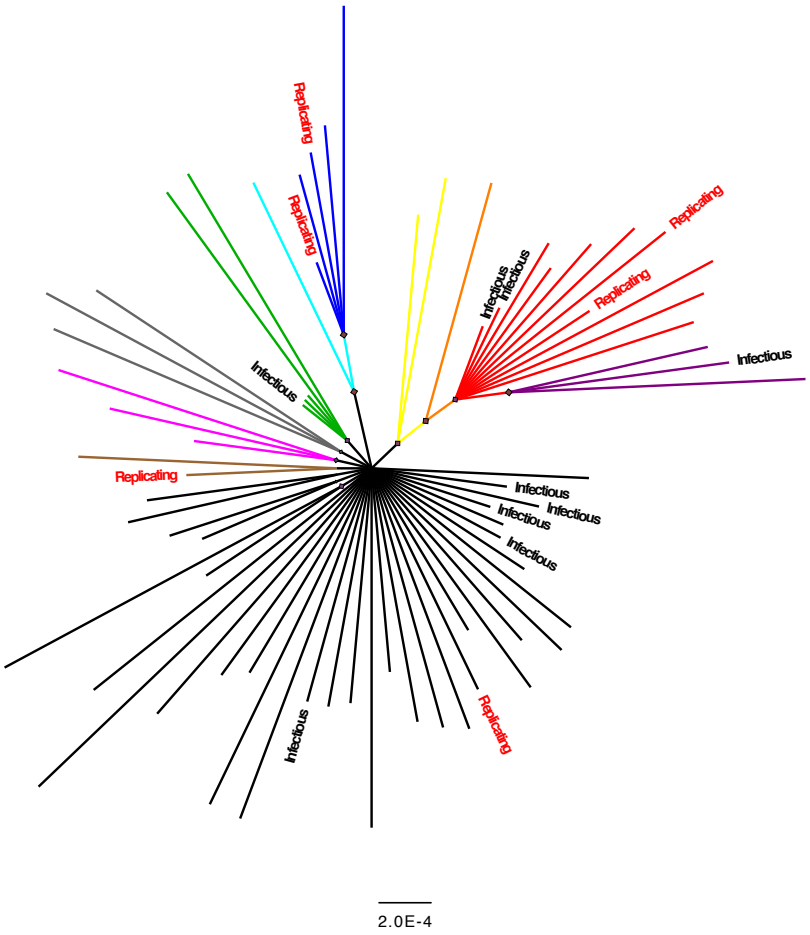


Fig. 2. Phylogenetic structure of cDNA clones. A) The structure of cDNA clone population is shown as a Bayesian inferred unrooted tree with functionality assigned to each "leaf". Different haplotypes are shown as colored clades.

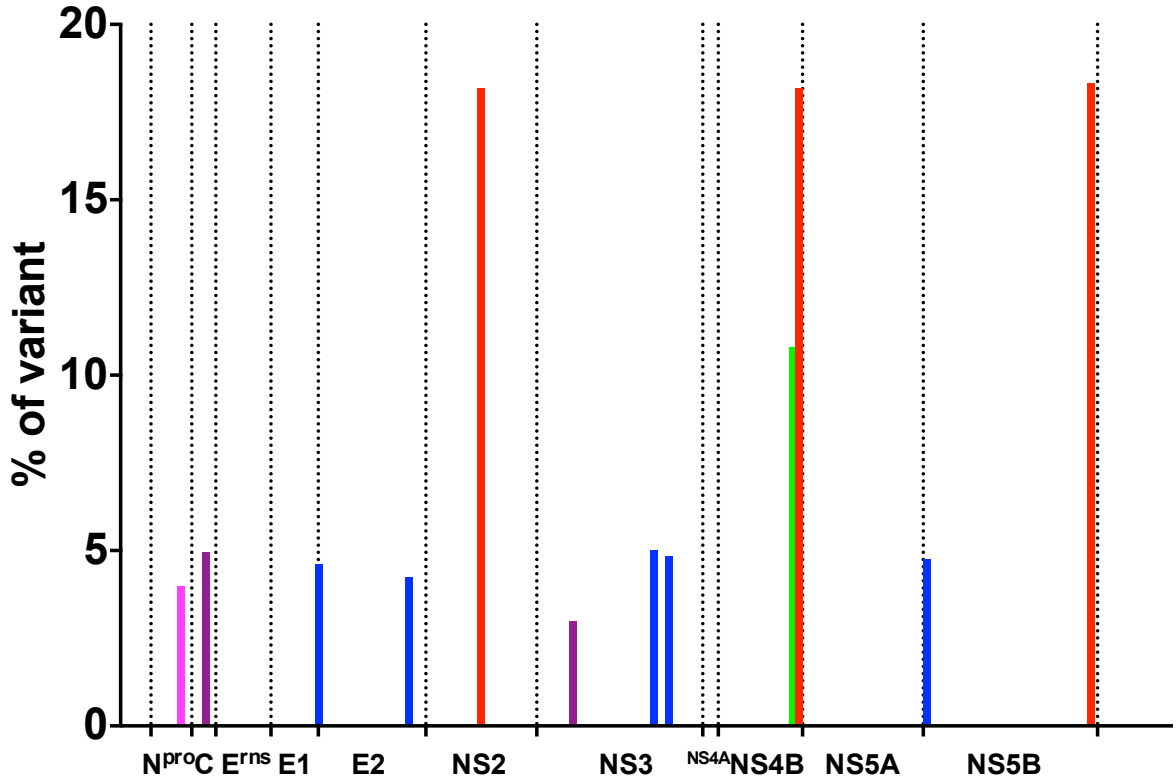


Fig. 3. Haplotype mutation frequency distribution. The plot shows the major haplotypes and mutation linkage from the NGS deep sequencing. The color coding corresponds to the respective haplotypes in fig. 2.

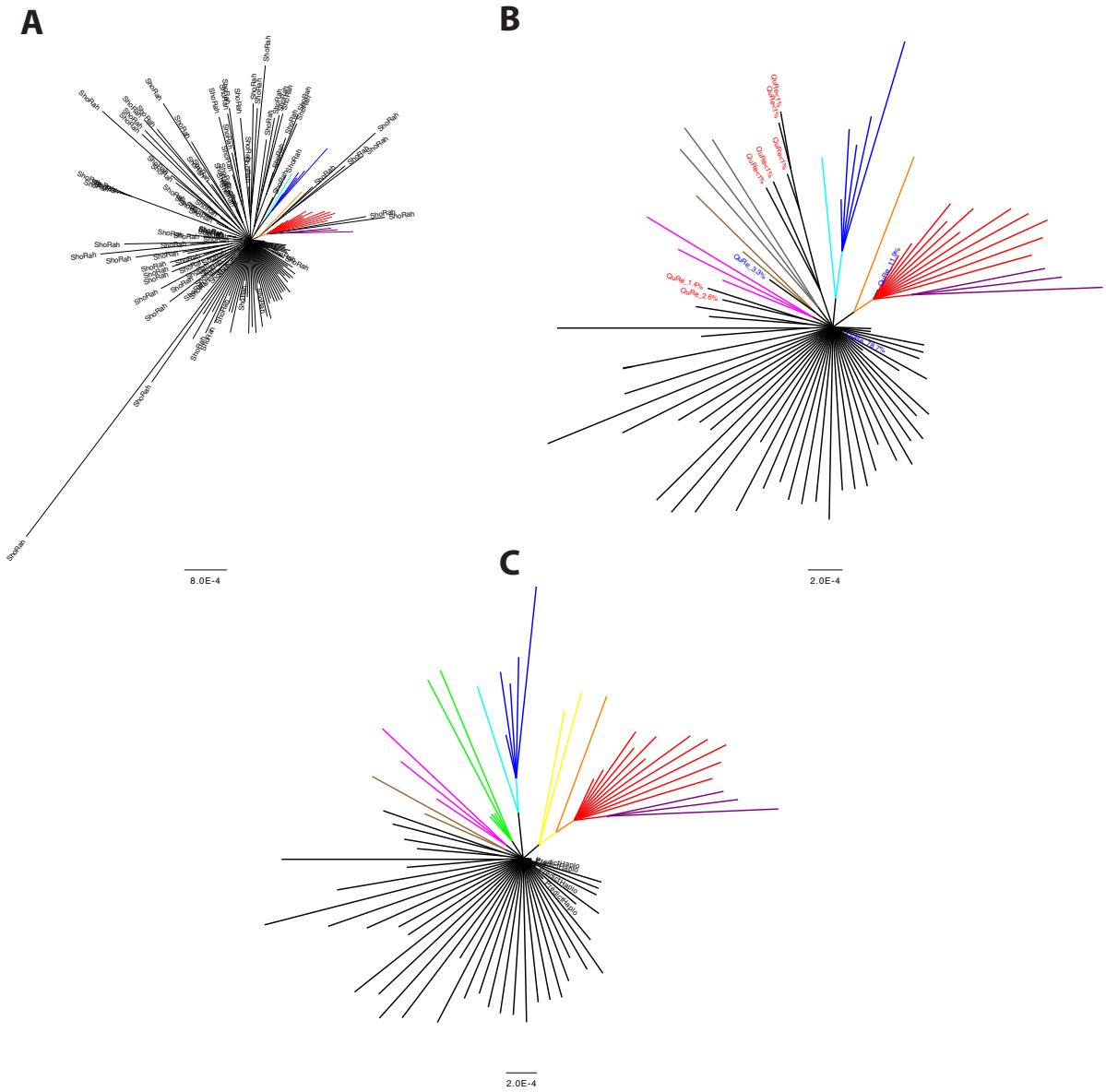


Fig. 4. Analysis of haplotype reconstruction of NGS data. The haplotype predictions are shown as a Bayesian inferred unrooted tree with functionality assigned to each "leaf". Different haplotypes are shown as colored clades. Each predicted haplotype is shown on the leafs. A) ShoRah B) QuRe C) PredictHaplo.

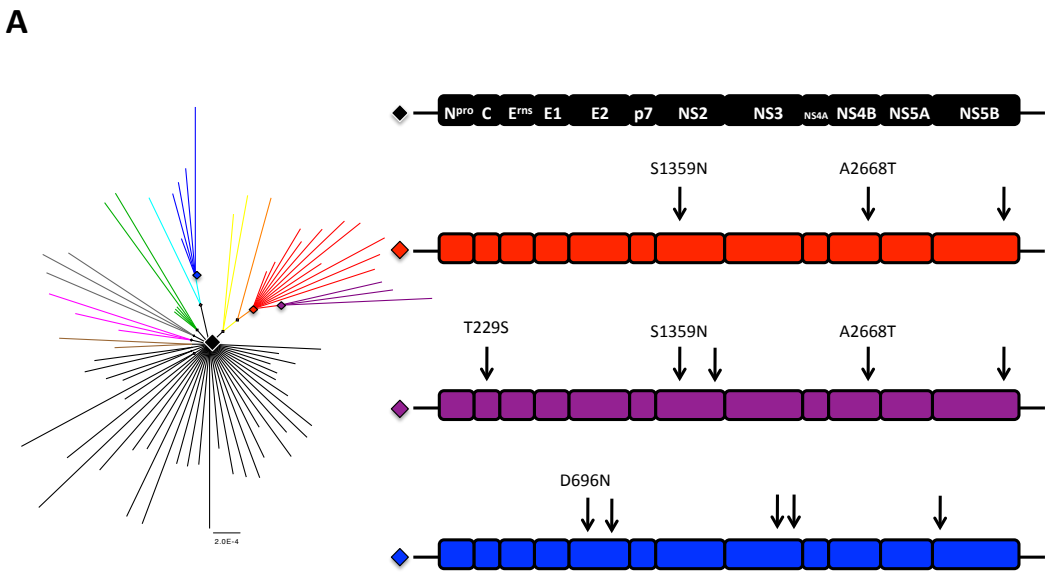


Fig. 5. Ancestral reconstruction of major internal nodes. A) Depiction of the ancestral reconstruction of the four major internal diamond shaped nodes. The deepest black node is identical to the reference sequence (GU233734) and the red, purple and blue predicted ancestral genomes are shown with the differences to the deepest ancestral node indicated. On the left the cDNA clone internal nodes reconstructed are shown. On the right side are the four major ancestral reconstructions differing by at least two SNPs. The arrows indicate mutations compared to the dominant black form. Mutations leading to aa changes are indicated. B) Reconstructed cDNAs corresponding to the two major internal nodes.

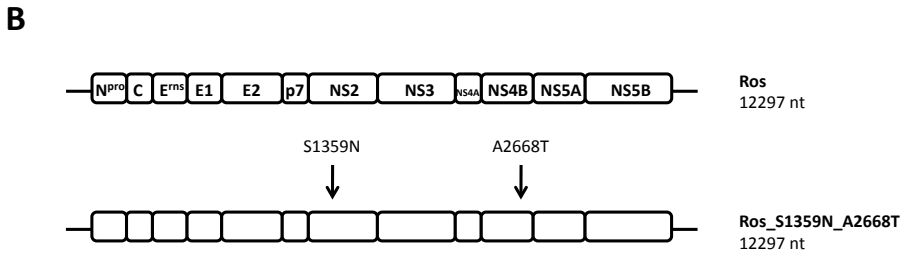


Fig. 5. Ancestral reconstruction of major internal nodes. A) Depiction of the ancestral reconstruction of the four major internal diamond shaped nodes. The deepest black node is identical to the reference sequence (GU233734) and the red, purple and blue predicted ancestral genomes are shown with the differences to the deepest ancestral node indicated. On the left the cDNA clone internal nodes reconstructed are shown. On the right side are the four major ancestral reconstructions differing by at least two SNPs. The arrows indicate mutations compared to the dominant black form. Mutations leading to aa changes are indicated. B) Reconstructed cDNAs corresponding to the two major internal nodes.

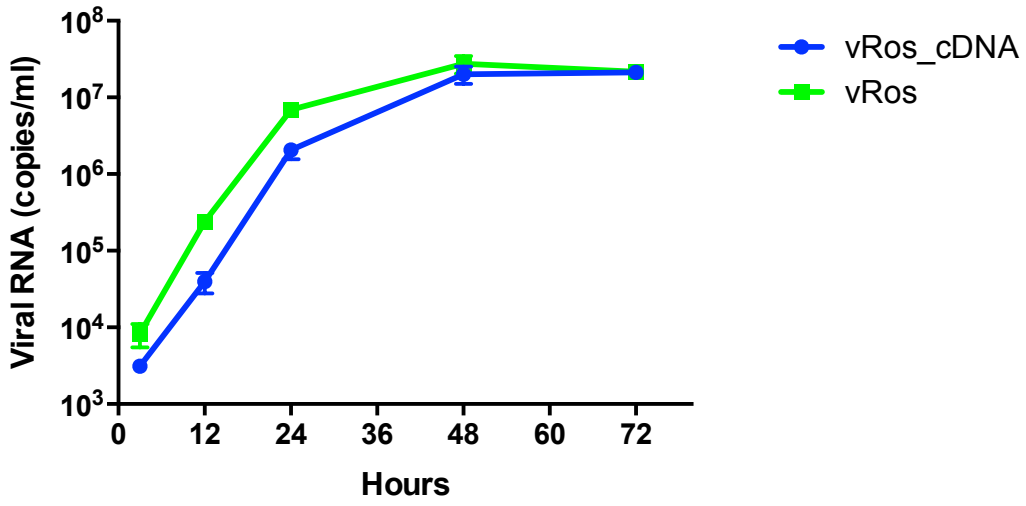


Fig. 6. Growth kinetics of the deepest ancestral node. Growth curves of viruses in PK-15 cells were measured using RT-qPCR assays (viral RNA copies/ml) at 3, 12, 24, 48, and 72 hours after infection. Means ± s.d. are shown for biological replicates (n = 3).

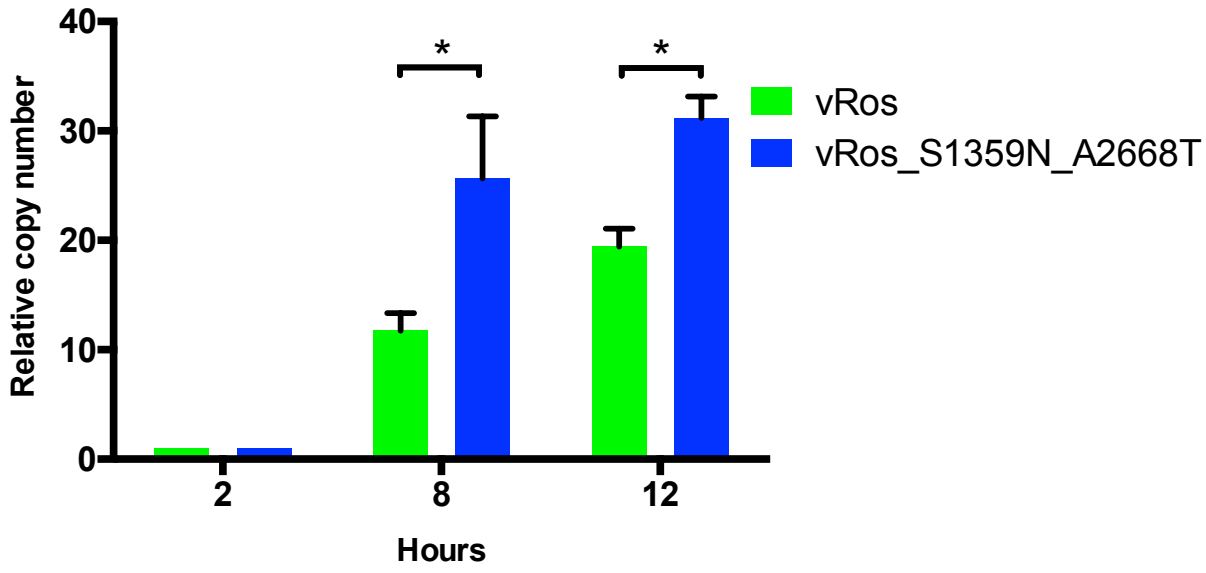


Fig. 7. Replication kinetics of both reconstructed nodes. Replication assay of viruses in PK-15 cells were measured using RT-qPCR relative to 2 hour measurement at 8 and 12 hours after infection. Means ± s.d. are shown for biological replicates (n = 3). The T-test was applied to determine significance differences between the clones at 8h (p=0.003) and 12h (p=0.008).

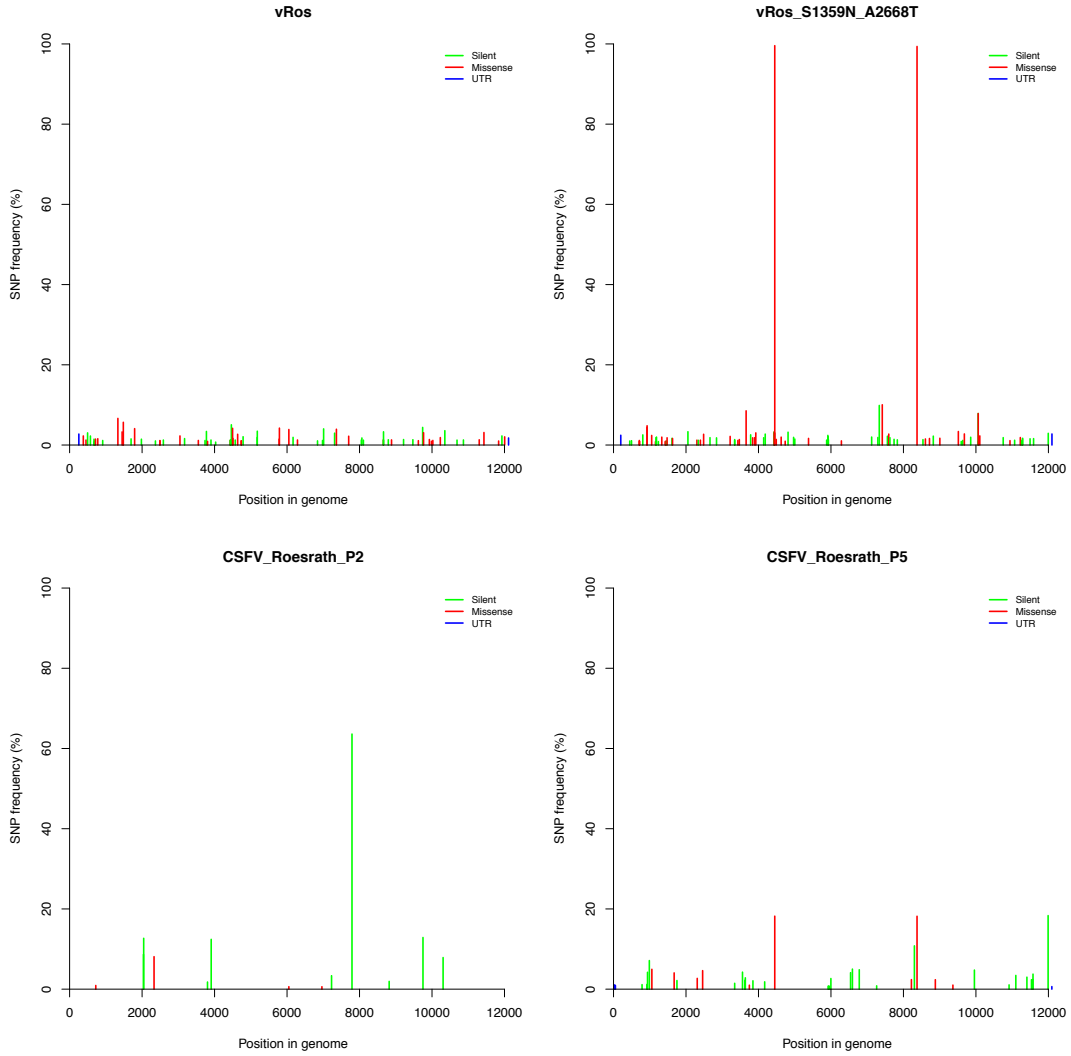


Fig. 8. Deep sequencing of virus inoculums for *in vivo* testing. The histograms depict SNP frequency on the y-axis and genome position on the x-axis. The blue, green and red color indicates SNPs grouped as silent, missense or untranslated region (UTR), respectively.

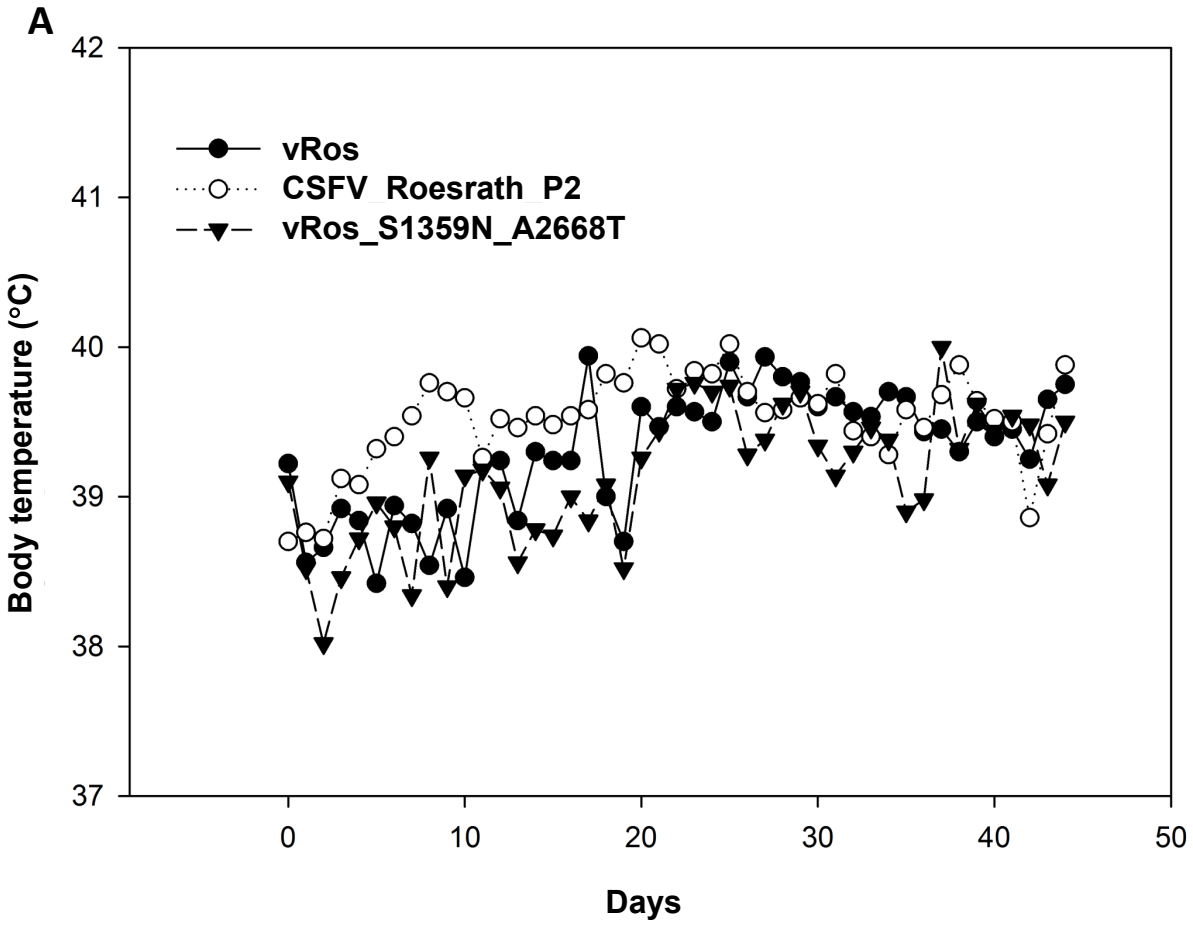


Fig. 9. Properties of reconstructed viruses in pigs. A) Body temperature during the course of infection. B) Level of viral RNA in the blood measured by RT-qPCR (viral RNA copies/ml). C) Level of viral RNA in oral swab samples measured by RT-qPCR (viral RNA copies/ml). Means ± s.d. (n = 5) are shown.

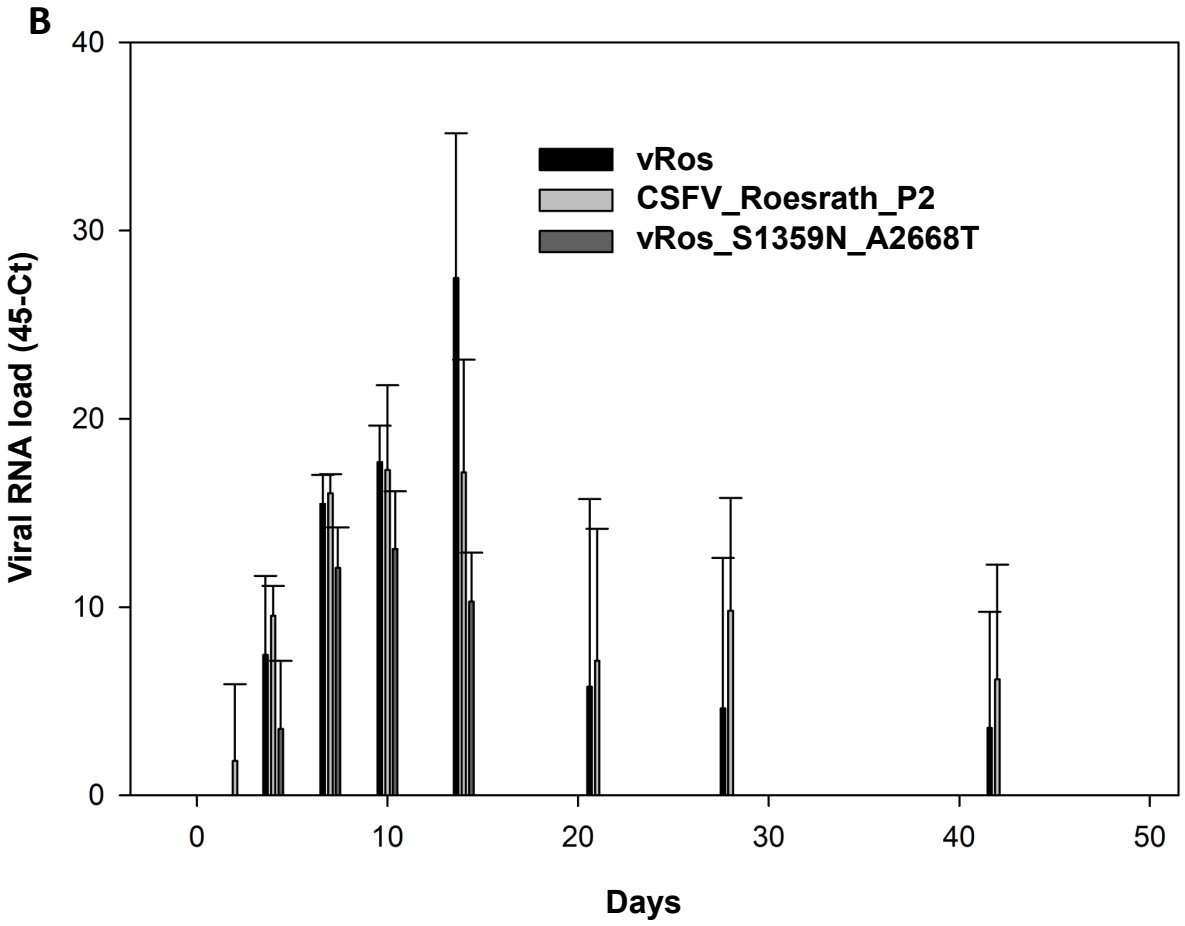


Fig. 9. Properties of reconstructed viruses in pigs. A) Body temperature during the course of infection. B) Level of viral RNA in the blood measured by RT-qPCR (viral RNA copies/ml). C) Level of viral RNA in oral swab samples measured by RT-qPCR (viral RNA copies/ml). Means ± s.d. (n = 5) are shown.

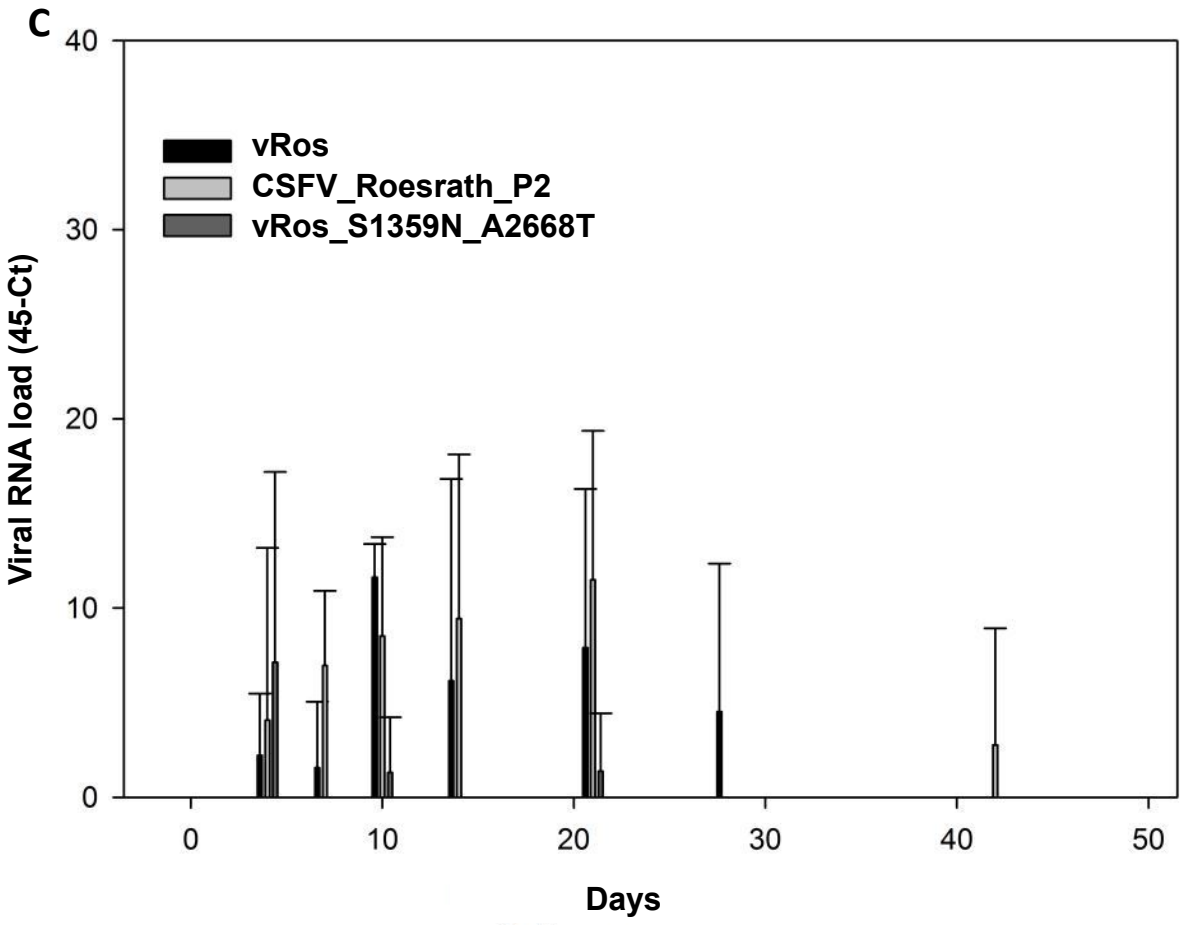


Fig. 9. Properties of reconstructed viruses in pigs. A) Body temperature during the course of infection. B) Level of viral RNA in the blood measured by RT-qPCR (viral RNA copies/ml). C) Level of viral RNA in oral swab samples measured by RT-qPCR (viral RNA copies/ml). Means ± s.d. (n = 5) are shown.

Manuscript 5

Classical swine fever virus adaptive response to vaccination: early signs of haplotype tropism

Classical swine fever virus adaptive response to vaccination: early signs of haplotype tropism

Ulrik Fahnøe[a,b], Anders Gorm Pedersen[b], Richard J Orton[c,d], Dirk Höper[e], Martin Beer[e], Thomas Bruun Rasmussen[a,#]

[a]*DTU National Veterinary Institute, Technical University of Denmark, Lindholm, DK-4771 Kalvehave, Denmark*

[b]*Center for Biological Sequence Analysis, DTU Systems Biology, Technical University of Denmark, Denmark*

[c]*Institute of Biodiversity, Animal Health, and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, UK*

[d]*MRC – University of Glasgow Centre for Virus Research, Institute of Infection, Inflammation and Immunity, College of Medical, Veterinary and Life Sciences, University of Glasgow, UK*

[e]*Institute of Diagnostic Virology, Friedrich-Loeffler-Institut, Greifswald-Insel Riems, Germany*

#Corresponding author:

Mailing address: DTU National Veterinary Institute, Technical University of Denmark, Lindholm, DK-4771 Kalvehave, Denmark

Phone: +45 3588 7850. Fax: +45 3588 7850. E-mail: tbrur@vet.dtu.dk

Highlights:

- Detailed SNP analysis allowed the recognition of patterns in vaccinated animals not seen in the control animals.
- SNP analysis indicated a haplotype tropism effect in pigs infected with CSFV strain "Koslov" displaying higher levels of the SNP S763L in the tonsils compared to the blood.
- dN/dS analysis revealed higher positive selection of the challenge virus in the immunised pigs.

Abstract

Next Generation Sequencing (NGS) can be applied to unravel molecular adaptation of RNA viruses such as Classical Swine Fever Virus (CSFV). Here we investigate adaptive responses of the highly virulent CSFV strain "Koslov" by analysis of NGS data obtained from immunised pigs. NGS data derived from the challenge virus in immunised pigs was compared to NGS data obtained from pigs without immunisation. Comparison of low frequency single-nucleotide polymorphisms (SNPs) revealed differences in the challenge virus populations between the immunised and the control animals. The SNP profile observed in the control pigs was very similar to the distribution seen in the virus inoculum used for the challenge infection. In contrast, the SNP profile in the immunised pigs displayed several distinct changes compared to the virus inoculum. This included the emergence of unique non-synonymous SNPs not present at a detectable level in the virus inoculum, indicating strong selective pressure imposed by the immunisation. In agreement with this, dN/dS analysis confirmed a significantly higher level of positive selection in the immunised pigs.

Introduction

RNA viruses have some of the highest mutation rates known in nature. This allows them to adapt quickly to selective pressures such as host immune responses (Drake 1993). Classical swine fever virus (CSFV), an RNA virus belonging to the genus *Pestivirus* within the *Flaviviridae* family, also has this inherent feature. CSFV is the causative agent of classical swine fever (CSF) and exists as multiple genotypes with varying phenotypes ranging from high virulent to low virulent types (Floegel-Niesmann et al. 2009). Studies in CSFV have revealed that highly virulent viruses have higher diversity compared to viruses of lower virulence (Töpfer et al. 2013). Whether this high diversity is necessary for high virulence is not fully understood. The ability of CSFV to adapt quickly during virus replication has been observed in modified live vaccine mutants where key amino acids revert to their parental state after a few passages in cell cultures (Rasmussen et al. 2013). Studies of CSFV adaptation in vivo of another live attenuated vaccine strain (GPE^-) was also found to revert key motifs after 11 passages in pigs resulting in a more virulent form (Tamura et al. 2012). Furthermore, adaptation events towards higher virulence also occurred within animals infected with a mutant form of the highly virulent CSFV strain "Koslov" (Fahnøe et al. 2014). Studies of viral adaptation under high selective pressure (such as antiviral treatment, neutralising antibodies or vaccination) have the potential to show adaptive escape variants being selected for. Examples of this include in vivo and in vitro studies of Hepatitis C virus (Thiel et al. 2002; Serre et al. 2013) and CSFV (Leifer et al. 2012).

Vaccination studies typically focus on the efficacy and safety of the CSF vaccine candidates (Beer et al. 2007). However, vaccinated individuals often show low-level transient viral RNA loads after the challenge infection (Rasmussen et al. 2007; Blome et al. 2012; Blome et al. 2014; Gabriel et al. 2012; Eble et al. 2013). This indicates replication of the challenge virus under strong selective pressure imposed by the immune system. Detailed analysis of the virus subpopulations during this transient viremic period may reveal adaptive changes at the molecular level. Further exploration of adaptation events in immunised animals will facilitate a better understanding of the adaptive potential of the challenge virus and thereby the protective capabilities of the vaccine candidates. Next generation sequencing (NGS) technologies has made it possible to study the evolution of viral populations in detail. In particular, use of NGS allows identification of low frequency single nucleotide polymorphisms

(SNPs) in virus populations, something that has previously been possible only by end-point limiting dilution or extensive cDNA cloning.

In this study, NGS sequencing of CSFV under strong selective pressure was performed on samples obtained from pigs immunised with two different CSF vaccine candidates (vR26 and vR26_E2gif) and challenge infected with the highly virulent CSFV strain "Koslov". The deep sequencing allowed detailed SNP analyses of challenge virus populations from both immunised and control animals. A modified sample preparation scheme made NGS of the challenge virus possible from samples with low viral loads and NGS error correction enabled in depth comparison of low frequency SNPs.

Results

Transient viral RNA loads of challenge virus in immunised pigs

Pigs immunised with vR26 and vR26_E2gif (Rasmussen et al. 2013) were protected from lethal challenge and only transient fever was observed in all immunised animals after challenge infection with the highly virulent CSFV strain "Koslov" (Fig. 1a)(Rasmussen et al., unpublished). The highest temperatures were seen for pigs immunised with vR26_E2gif. All immunised pigs seroconverted within 21 days after vaccination (Rasmussen et al., unpublished). After challenge infection low levels of viral RNA were observed between post infection day (PID) 3 and PID10 in blood samples from vR26_E2gif, whereas vR26 immunised pigs showed low viral RNA levels between PID3 and PID7 (Fig. 1c). Transmission to sentinel pigs was not observed. All sentinel pigs remained negative in RT-qPCR throughout the experiment and did not seroconvert (Rasmussen et al., unpublished).

All pigs in the challenge infected control group developed high fever shortly after challenge infection (Fig. 1a) and were euthanized due to severe clinical signs of CSF within 6 days (Fig. 1b). High viral loads were observed in both blood (Fig. 1c) nasal swabs and tonsils (Rasmussen et al., unpublished).

NGS data from immunised and challenge infected animals

To investigate adaptation events in the challenge virus populations, SNP analysis was performed on NGS data from the virus inoculum and from the immunised and control animals (Supplementary table 1). First, the Koslov virus population used for the challenge infection was sequenced from a full-length cDNA and furthermore also sequenced directly from the RNA (Fig. 2a,b). Sequencing of the cDNA revealed the consensus to be 100% identical to the published reference sequence for the CSFV strain "Koslov" (Genbank HM237795). Furthermore, variant calling shows that this consensus sequence is representative of a heterogeneous viral population (Fig. 2a) with 10 SNPs having approximately 40% SNP frequencies. One of these is a missense SNP (C2661T, amino acid change S763L) in the E2 protein, whereas the other 9 are silent SNPs (C2134T, G3205A, T4150C, C4612G, T4750C, G5101A, T9940C, A10669G and G11374C). All other (149 SNPs) had frequencies between 1 and 10% with the majority of these corresponding to silent mutations. Direct sequencing of the inoculum RNA confirmed the results by showing similar SNP profile as obtained for the RT-PCR product (Fig. 2b).

Our full-length RT-PCR amplification approach works well on samples with high viral RNA loads and low amounts of host RNA. The serum samples (ct value below 20) from the end-point of the control pigs (PID6) proved to be ideal for full-length RT-PCR as well as direct RNA sequencing. From the control group, we sequenced the serum samples (PID6) from each pig by RNA sequencing (Fig. 2d). RNA extracted from serum at PID6 was used for RNA sequencing with 99 % of the reads mapped to the reference sequence (data not shown), indicating that high titre samples from non-cellular samples were feasible. However, low titres samples from serum (Ct values higher than 25) could not be efficiently amplified by full-length RT-PCR and could not be sequenced using this approach. Therefore, a modified sample preparation scheme was developed to amplify the cDNA in two overlapping fragments each of approximately 6000 bp. The RT step was modified with an extra internal specific primer annealing in the middle of the genome allowing the 5′ end part of the genome to be amplified with similar efficiency as the 3′ part (data not shown). This modified protocol allowed us to sequence serum from the control group at PID3 (Fig. 2c).

Importantly, this modified protocol also allowed sequencing of serum samples with low viral loads from immunised pigs. From the immunised pigs, we sequenced a total of four serum samples (3 from vR26_E2gif and 1 from vR26) obtained at PID5 from individual immunised pigs (Fig. 2 g, h, i, j).

Finally, tonsil samples from the control animals obtained at PID6 were sequenced by full-length RT-PCR products (Fig. 2f). In order to get adequate amounts of RT-PCR product from tissues (in this case tonsils) that contain excessive amounts of host RNA a modified RNA extraction was applied. After RNA extraction five eluates were full-length RT-PCR amplified. Strongest and clearest bands were seen for eluates three to five (table 1). The Ct values determined by RT-qPCR of the individual RNA eluates indicates that it is not the amount of viral RNA that is important, but rather the proportion of intact viral RNA (relative to the total amount of RNA) and the integrity.

Comparison analysis in immunised and challenge infected animals

SNP frequency analysis was performed on the NGS data from both the immunised and the control animals. This analysis revealed altered frequencies of SNPs in the two groups of animals (Fig. 3). In general, the number of SNPs with 1-10% frequencies was reduced for the immunised pigs compared to the control pigs. Above 10%, the immunised pigs have 6 SNPs close to fixation in the populations, whereas the control pigs have 4 SNPs at 50-60% in serum.

Further SNP analysis was done in order to look for pattern in the SNP calls between each group of animals and the virus inoculum. Figure 4a shows the SNP comparison of serum samples depicted as a Venn-diagram. The control pigs generally have the same set of SNPs as the inoculum at PID3 in serum. Similar patterns were detected for blood, tonsil and serum at PID6 (data not shown), and almost all SNPs of the virus inoculum were still present in the control pigs. However, serum samples obtained from immunised pigs (PID5) seemed to maintain less of the inoculum variation and were diversified in different directions.

The individual immunised pigs had diversified not only in comparison to the virus inoculum but also between each other (fig. 4a). This diversification could be due to high

selective pressure caused by the immunisation. Thus, the sequence data obtained from immunised and challenge infected pigs can tell us about unique missense mutations that may correspond to adaptations. Unique missense SNPs from each of the four immunised animals are shown in figure 5. Only low frequency unique missense SNPs were detected (1-8%). As can be seen each sample do not show the same number of unique SNPs and the changes appear to be scattered along the genome. Nevertheless, some regions seem more prevalent and are seen in more than one sample. This includes the structural proteins E^{rns}, E1 and E2 and the non-structural proteins NS5A and NS5B.

In order to study these patterns of shared SNPs more closely, intersections of the individual pig samples were performed with SNPs shared at three out of four pigs and two out of three pigs for the immunised and the control groups, respectively. The comparison of the intersections clearly showed a reduced amount of shared population diversity among the immunised animals and also a large proportion of SNPs that were not shared with either the inoculum or the serum from control pigs pointing to host immune response adaptation (Fig. 4b). However, the Venn-diagrams do not reveal the individual changes in SNP frequency shared by the different groups, which could also play an important role in the virus adaptation. To address both emergent and frequency altered SNPs compared to the inoculum, we constructed plots showing the average change in SNP frequencies (Δ%) for both control and immunised pigs (Fig. 6a and b). The plots differ with the immunised group having a new consensus sequence at six positions (C2661T, G3205A, C4612G, T4750C, T9940C and A10669G), with especially the C2661T missense mutation S763L being close to fixation and several missense SNPs occurring in between 1-10% of the populations (table 2). Additionally, some of the missense SNPs were not found in the inoculum above the detection level suggesting mutation followed by positive selection (table 2). We observed new mutations in the N^{pro} and at the border between P7 and NS2 that might have some significance. This was not observed for the control pig serum samples where the S763L mutation is reduced in frequency, which was also the case at PID6 in serum (data not shown). Instead 4 silent consensus changes between 50-60% were observed (C2134T, T4150C, G5101A and G11374C). However, similar analysis of tonsil and blood samples from the control pigs did not have the same tendency. Instead the blood samples closely resembled the inoculum with no major changes (Fig. 6c). The tonsil samples displayed Δ% SNP frequencies increase by 17% of

the S763L mutation leading to a change in consensus together with the same 5 silent positions as the immunised pigs though without fixation (Fig. 6d and table 2). Several of the low frequency missense SNPs otherwise only found in serum from immunised animals were detected in the tonsils from control pigs though at lower frequencies (table 2). These results point to two major haplotypes dominating the virus inoculum, one with 4 silent SNPs (C2134T, T4150C, G5101A and G11374C), the other containing 6 SNPs (C2661T, G3205A, C4612G, T4750C, T9940C and A10669G) of which the C2661T translated to S763L mutation is the only missense. The differences in SNP profiles between the different types of samples in the control animals indicate haplotype tropism where the haplotype with S763L mutation as the dominating variant in the tonsil samples and a status quo is seen in the blood samples, while serum samples seem to favour the S763 variant.

Positive selection was observed in immunised pigs

In order to further investigate selective pressure dN/dS analysis was carried out on the individual NGS data from each sample. The mean dN/dS for the entire polyprotein is plotted in figure 7. We observed the highest dN/dS ratios for the inoculum, the immunised animals and the tonsil samples compared to the blood and serum samples from the control pigs. In addition, there was a significant difference between immunised and the control animals in serum (fig. 7b). However, no significant difference was observed between the tonsil from the control group and the serum samples from the immunised animals. We also observed a significant difference between the viral RNA in the tonsils of control animals compared to serum and blood (Fig. 7c). These trends could be due to both purifying selection working on the blood and serum of the control animals, and the virus trying to escape the high selection pressure in the immunised animals leading to a slightly higher ratio although still purifying ($dN/dS < 1$).

Discussion

This study describes an investigation of challenge virus populations under the selective pressure imposed by immunisation using NGS data. Detailed SNP and dN/dS analysis allowed

us to discover patterns in immunised animals not seen in the mock-immunised animals. In addition, we detected several amino acid residues under positive selection and an apparent haplotype tropism effect in the control animals. We suggest that these analyses would benefit virus vaccine studies by allowing identification of adaptive mutations and sites important for potential immune escape of the challenge virus. The knowledge gathered here may be useful for production of better live attenuated vaccines and lead to better understanding of the adaptive potential of a virus.

A modified RT-PCR protocol allowed the deep sequencing of the Koslov challenge strain at low titers (ct values 28-33) as two overlapping fragments. In addition, a modified full-length RT-PCR protocol for tissues was developed, and analysis revealed that host RNA inhibited the RT-PCR at high concentrations with RNA elutes 3 to 5 giving the best results.

The main focus of this study was the analysis of the molecular evolution of the viral populations based on NGS data produced using the methods described above. The virus inoculum used for the challenge infection was shown to be a heterogeneous population with two main haplotypes and many low frequency SNPs. This may be the result of balancing selection maintaining variation within the population, as has previously been observed for other highly virulent CSFV populations (Töpfer et al. 2013). However, whether this heterogeneity is in itself necessary for the high virulence is unclear (Fahnøe et al. 2014). SNP analysis also revealed a decline in low frequency variants for immunised pigs compared to controls along with several changes in the consensus sequence. A decrease was observed in both the number of SNPs in the immunised pigs, as well as in the number of SNPs shared between individual pigs and shared with the inoculum. Comparative analysis showed clear differences between immunised and control pigs. For instance, the S763L SNP was almost at fixation in all immunised pigs, which also possessed several low frequency missense SNPs not found in the inoculum or in serum samples from controls. Virus populations isolated from tonsils were more similar to immunised serum samples although the SNPs were present at lower frequencies. Taken together this suggests that the S763L haplotype is replicating more efficiently in the tonsils and that this haplotype may be partially protected from the immune response, resulting in the near fixation observed in serum from immunised pigs. However, S763 is still present above 50% in all controls, in both blood and serum, indicating that this haplotype is efficiently replicating elsewhere in the pig. It seems relevant that amino acid 763 is situated in a putative epitope in the E2 surface protein of the virus (Chang et al. 2012),

although further studies into this epitope are needed to firmly understand these phenomena. The missense mutations present at low frequencies, which are only found in immunised pigs and some tonsil samples, could be positively selected sites. Several of the missense SNPs in the N^{pro}, E1, E2 and p7 could play a role in viral immune escape as they are either in structural proteins that could interact with the immune system or in proteins known to inhibit the anti-viral response.

dN/dS analysis indicated that challenge virus in serum from immunised pigs displayed a slightly higher level of positive selection compared to virus in control pigs. This is consistent with an increased selective pressure being exerted by the immune system in the immunised pigs. In addition, within the control animals virus isolated from tonsils displayed significantly higher levels of positive selection compared to both blood and serum from the same animals.

Materials and Methods

Vaccine and virus

The 12^{th} passages (in SFT-R cells) of the recombinant C-strain vaccine vR26 and the chimeric derivative vR26_E2gif (Rasmussen et al. 2013) were used for the immunisation of pigs. Blood from a CSFV strain "Koslov" (CSFV/1.1/dp/CSF0382/XXXX/Koslov) infected pig was used as virus inoculum for the challenge infection (Blome et al. 2014).

Immunisation and challenge infection of animals

An immunisation/challenge experiment to assess the marker vaccine properties of vR26_E2gif, in comparison to vR26, was previously performed (Rasmussen et al. unpublished). In brief, two groups of 6 pigs (6 weeks old) in separate pens were immunised intramuscularly with either vR26_E2gif or vR26 (Rasmussen et al. 2013). Each of these pens also housed three sentinel pigs. Three animals, housed in a separate pen, were mock-immunised with cell culture medium and served as challenge infection controls. Four weeks later (PID0) all pigs, except for the sentinels, were inoculated with highly virulent CSFV strain

"Koslov" (2 × 10^6 $TCID_{50}$). All pigs were observed for clinical signs typical for CSF, body temperatures were recorded, blood samples and nasal swabs were examined for viral RNA load by qRT-PCR and serum samples were tested for anti-CSFV antibodies by ELISA (Rasmussen et al. 2007) (Nielsen et al. 2010).

RNA extraction and NGS

RNA was extracted from selected samples using a modified Trizol/RNeasy protocol as previously described (Rasmussen et al. 2010). Briefly, RNA was extracted from virus inoculum, blood, serum and tonsils by Trizol LS (Invitrogen), the aqueous phase containing the RNA was washed on an RNeasy spin column (Qiagen) and RNA was eluted in nuclease free H_2O. For RNA extraction from tonsils, the tissue was cut into small pieces with a sterile pair of scissors. 100 mg of tissue was added 750 µl Trizol LS and 250 µl H_2O in a 1.5 ml tube. One steel ball was placed in each tube and was run 2 times on a Tissuelyzer (Qiagen) at 25 Hz for 1 min. The rest of the extraction was performed as above.. The RNA from virus inoculum, blood and tonsils (with Ct values above 20) were amplified by full-length RT-PCR as described in (Fahnøe et al. 2014), whereas the RNA from pigs with low viral loads (with Ct values above 28) were amplified in two fragments of approximately the same size using a modified protocol. This modified protocol makes use of a RT step primed by two specific primers, and subsequently amplifying the cDNA with two PCR reactions to cover the entire genome (5' end: CSF-Kos_Not1-T7-1-59 5'TCT ATA TGC GGC CGC TAA TAC GAC TCA CTA TAG TAT ACG AGG TTA GTT CAT TCT CGT ATG CAT GAT TGG ACA AAT CAA AAT TTC AAT TTG G 3' CSF-Kos-6176-R 5' CTG GTG TTG CGG TCA TGG CTA CTA C 3') (3' end CSF-Kos-5981-F 5'GGG GAG ATG AAA GAA GGG GAC ATG 3' CSF-kos_12313aR 5' GGG CCG TTA GGA AAT TAC CTT AGT CCA ACT GT 3'). The two fragments were pooled in equal amounts (2 x 250 ng) and sequenced. RNA extracted from blood by MagnaPure extraction (Roche), used for the RT-qPCR, was also amplified by this modified protocol using the primers mentioned above. NGS was performed on both the RT-PCR products and on RNA using the Ion PGM and the FLX platforms (Supplementary table 1).

Sequence Data Analysis

FastQC (Andrews 2010) was applied for pre-trimming evaluation of the Fastq files containing the raw sequence reads. Trimming and primer removal was performed by cut-adapt and prinseq-lite (Schmieder and Edwards 2011). Fastq files were error corrected by RC454 (Henn et al. 2012) with the CSFV strain "Koslov" nucleotide sequence (GenBank HM237795) as reference. We performed the *de novo* assembly by Newbler 2.6 (Roche software). Each error corrected fastq file was aligned by BWA-MEM (Li 2013) algorithm or Mosaik (Lee et al. 2014). Subsequently, the libraries were post processed by samtools (Li et al. 2009) and SNPs were called by lofreq (Wilm et al. 2012). Downstream SNP effect analysis was performed by snpEff (Cingolani et al. 2012). SNP plots were done in R as were the Venn-diagrams using the venneuler module and VCFtools (Danecek et al. 2011). The Vcf-compare module of the VCFtools provided the numbers for the Venn-diagrams. VCFtools were also used to intersect VCF files from different samples.

dN/dS analysis

In order to estimate the non-synonymous to synonymous ratio dN/dS in the NGS datasets, we use the approach proposed by (Morelli et al. 2013) which is based on (Nei and Gojobori 1986). We first computed the expected number of synonymous (s_i) and non-synonymous (n_i) sites for each codon *i* in the viral genome open reading frame (ORF). Then, for each codon *i* we counted the observed number of synonymous (s_{di}) and non-synonymous (n_{di}) mutations in all reads that fully covered codon *i*. The proportion of non-synonymous differences (p_n) in the entre NGS dataset across the entire ORF is then calculated with the following formula:

$$p_n = \sum_{i=1}^{r} \frac{1}{c_i} \sum_{j=1}^{c_i} \frac{n_{dij}}{n_i}$$

Briefly, *i* is the codon number within the ORF of length *r* codons, c_i is the read coverage at codon *i* (only reads that fully cover codon *i* are considered), n_{dij} is the number of observed

non-synonymous mutations at codon *i* in read *j*, and n_i is the expected number of non-synonymous mutations at codon *i*. The same formula is used to calculate the proportion of synonymous differences (p_s) in a similar fashion. The dN/dS ratio is then determined from p_n and p_s as described in (Nei and Gojobori 1986; Morelli et al. 2013).

Acknowledgements

This work was supported by the European project Epi-SEQ (research project supported under the 2nd Joint Call for Transnational Research Projects by EMIDA ERA-NET [FP7 project no. 219235]) and by the German Federal Ministry for Education and Research (BMBF, grant 01KI1016A).

References

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Beer M, Reimann I, Hoffmann B, Depner K. 2007. Novel marker vaccines against classical swine fever. Vaccine 25:5665-5670.

Blome S, Aebischer A, Lange E, Hofmann M, Leifer I, Loeffen W, Koenen F, Beer M. 2012. Comparative evaluation of live marker vaccine candidates "CP7_E2alf" and "flc11" along with C-strain "riems" after oral vaccination. Vet. Microbiol. 158:42-59.

Blome S, Gabriel C, Schmeiser S, Meyer D, Meindl-Böhmer A, Koenen F, Beer M. 2014. Efficacy of marker vaccine candidate CP7_E2alf against challenge with classical swine fever. Vet. Microbiol. 169:8-17

Chang CY, Huang CC, Deng MC, Huang YL, Lin YJ, Liu HM, Lin YL, Wang FI. 2012. Antigenic mimicking with cysteine-based cyclized peptides reveals a previously unknown antigenic determinant on E2 glycoprotein of classical swine fever virus. Virus Res. 163:190-196.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms,

SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80-92.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. (x co-authors. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

Drake JW. 1993. Rates of spontaneous mutation among RNA viruses. Proc. Natl. Acad. Sci. U. S. A. 90:4171-4175.

Eble PL, Geurts Y, Quak S, Moonen-Leusen HW, Blome S, Hofmann MA, Koenen F, Beer M, Loeffen WL. 2013. Efficacy of chimeric pestivirus vaccine candidates against classical swine fever: Protection and DIVA characteristics. Vet. Microbiol. 162:437-446.

Fahnøe U, Pedersen AG, Risager PC, Nielsen J, Belsham GJ, Höper D, Beer M, Rasmussen TB. 2014. Rescue of the highly virulent classical swine fever virus strain "koslov" from cloned cDNA and first insights into genome variations relevant for virulence. Virology 468-470C:379-387.

Floegel-Niesmann G, Blome S, Gerss-Dulmer H, Bunzenthal C, Moennig V. 2009. Virulence of classical swine fever virus isolates from europe and other areas during 1996 until 2007. Vet. Microbiol. 139:165-169.

Gabriel C, Blome S, Urniza A, Juanola S, Koenen F, Beer M. 2012. Towards licensing of CP7_E2alf as marker vaccine against classical swine fever-duration of immunity. Vaccine 30:2928-2936.

Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf CM, Ryan EM, Gnerre S et al.. 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 8:e1002529.

Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. 2014. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One 9:e90581.

Leifer I, Blome S, Blohm U, König P, Kuster H, Lange B, Beer M. 2012. Characterization of C-strain "riems" TAV-epitope escape variants obtained through selective antibody pressure in cell culture. Vet. Res. 43:33-9716-43-33.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079.

Li H. Aligning sequence reads, clone sequences and assembly

contigs with BWA-MEM. 2013. arXiv:1303.3997v2. http://arxiv.org/abs/1303.3997v2.

Morelli MJ, Wright CF, Knowles NJ, Juleff N, Paton DJ, King DP, Haydon DT. 2013. Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. Vet. Res. 44:12-9716-44-12.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418-426.

Nielsen J, Lohse L, Rasmussen TB, Uttenthal A. 2010. Classical swine fever in 6- and 11-week-old pigs: Haematological and immunological parameters are modulated in pigs with mild clinical disease. Vet. Immunol. Immunopathol. 138:159-173.

Rasmussen TB, Reimann I, Uttenthal A, Leifer I, Depner K, Schirrmeier H, Beer M. 2010. Generation of recombinant pestiviruses using a full genome amplification strategy. Vet. Microbiol. 142(1-2):13-7

Rasmussen TB, Risager PC, Fahnøe U, Friis MB, Belsham GJ, Höper D, Reimann I, Beer M. 2013. Efficient generation of recombinant RNA viruses using targeted recombination-mediated mutagenesis of bacterial artificial chromosomes containing full-length cDNA. BMC Genomics 14:819-2164-14-819.

Rasmussen TB, Uttenthal A, Reimann I, Nielsen J, Depner K, Beer M. 2007. Virulence, immunogenicity and vaccine properties of a novel chimeric pestivirus. J. Gen. Virol. 88:481-486.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863-864.

Serre SB, Krarup HB, Bukh J, Gottwein JM. 2013. Identification of alpha interferon-induced envelope mutations of hepatitis C virus in vitro associated with increased viral fitness and interferon resistance. J. Virol. 87:12776-12793.

Tamura T, Sakoda Y, Yoshino F, Nomura T, Yamamoto N, Sato Y, Okamatsu M, Ruggli N, Kida H. 2012. Selection of classical swine fever virus with enhanced pathogenicity reveals synergistic virulence determinants in E2 and NS4B. J. Virol. 86:8602-8613.

Thiel J, Peters T, Mas Marques A, Rosler B, Peter HH, Weiner SM. 2002. Kinetics of hepatitis C (HCV) viraemia and quasispecies during treatment of HCV associated cryoglobulinaemia with pulse cyclophosphamide. Ann. Rheum. Dis. 61:838-841.

Töpfer A, Höper D, Blome S, Beer M, Beerenwinkel N, Ruggli N, Leifer I. 2013. Sequencing approach to analyze the role of quasispecies for classical swine fever. Virology 438:14-19.

Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40:11189-11201.

Table 1. RNA extraction and RT-PCR amplification from tonsils

RNA (Eluate)	1	2	3	4	5
RNA (ng/µl)	2719	1700	208	19	2
Ct	19.9	20.8	19.7	21.2	22.5
RT-PCR Band intensity*	(+)	+	+++	+++	++

*Band intensity was visualized on a 1% agarose gel and number of + is the intensity score.

Table 2. Positive SNPs in immunised animals and compared to control samples

Protein	Nucleotide position	Inoculum (nt)	Variant (nt)	Immunised (serum) SNP frequency %	Immunised (serum) SNP frequency Δ %	Control (serum) SNP frequency %	Control (tonsils) SNP frequency %	SNP effect	Comments
N^{pro}	425	A	G	5.1	5.1	-	-	M18V	*
	547	A	G	6.1	6.1	-	6.4	-	*
	658	C	T	7.7	3.0	-	3.6	-	
	725	A	G	5.1	5.1	-	1.7	I118V	*
	814	C	T	3.1	0.6	1.3	1.5	-	
C	1057	T	C	8.9	5.9	2.5	3.8	-	
E^{rns}	1774	A	G	4.6	4.6	-	1.3	-	*
E1	2024	T	C	13.8	7.2	1.52	9.4	F551L	
	2395	C	T	6.6	2.6	0.7	5.8	-	
E2	2617	T	C	9.5	5.8	2.8	5.5	-	
	2650	T	C	3.9	3.9	-	1.5	-	*
	2661	C	T	89.3	47.8	18.6	58.9	S763L	
	2935	G	A	6.7	4.2	1.65	3.1	M854I	
	3205	G	A	87.7	43.6	21.7	60.1	-	
	3244	A	G	6.5	3.5	0.6	6.4	-	
	3310	G	A	7.4	7.4	-	-	-	*
p7	3765	A	G	9.1	9.1	-	5.6	K1131R	*
NS2-3	3782	G	A	8.1	5.1	0.7	5.1	G1137S	
	3853	G	A	6.8	3.8	1.1	3.2	-	
	3907	G	A	2.9	2.9	-	-	-	*
	3931	T	C	1.8	1.9	-	2.0	-	*
	3965	G	A	1.5	1.5	-	-	V1198M	*
	4168	A	G	3.3	3.3	-	2.4	-	*
	4612	C	G	87.7	44.2	23.0	59.0	-	
	4729	A	G	15.2	7.2	2.7	9.2	-	
	4750	T	C	88.3	46.2	22.1	59.6	-	
	4760	G	A	6.2	6.2	-	3.6	V1463I	*
	4828	T	C	6.8	6.8	1.45	4.7	-	
	5248	A	G	6.0	2.2	3.0	3.4	-	
	5389	C	T	9.2	5.7	0.7	3.7	-	
	5416	T	C	15.4	9.5	1.9	9.3	-	
	6028	G	A	1.6	1.6	-	-	-	*
	7006	G	A	5.7	4.0	0.6	2.9	-	
	7021	T	C	5.1	1.9	0.6	5.7	-	
	7099	A	G	16.2	9.0	3.3	9.7	-	
NS4A	7304	T	C	6.9	3.4	0.7	1.4	-	
	7318	A	T	1.8	0.1	0.6	2.2	-	
NS4B	7696	T	C	7.2	4.5	2.3	4.0	-	
	7888	T	C	5.6	5.6	1	2.9	-	
	8341	C	T	9.5	5.9	1.4	5.0	-	
NS5A	8500	C	T	6.0	4.2	0.6	3.4	-	
	8878	G	A	9.7	5.8	1.2	3.7	-	

	9241	A	T	8.3	4.7	1.9	3.4	-	
	9396	A	G	7.8	2.5	9.2	9.9	K3008R	
	9871	G	A	5.8	5.9	1.4	3.6	-	
	9901	T	C	1.6	1.6	-	-	-	*
NS5B	9940	T	C	85.7	43.9	17.6	59.2	-	
	10259	T	C	4.8	4.8	-	-	-	*
	10402	C	T	6.9	4.3	1.3	3.1	-	
	10669	A	G	85.3	42.6	18.6	57.7	-	
	11242	T	C	5.7	5.7	-	-	-	*
	11452	C	T	5.4	2.3	-	6.1	-	
	11849	C	T	7.4	5.0	2.3	3.5	-	
	11872	A	G	6.5	3.5	-	4.0	-	
3' UTR	12083	C	T	6.3	3.3	-	1.4	-	

SNP frequencies are depicted as average within each group. Δ % is the average SNP frequency compared to the inoculum frequency. * SNPs in immunised only and not in controls and in virus inoculum.

Supplementary Table 1. Samples sequenced, NGS platform and data analysis.

Group	Material sequenced	Pig no.[1]	PID	RT-PCR[2]	NGS input	NGS platform	Error-correction	SNP caller
Inoculum	Blood	NA	0	-	RNA	FLX	RC454	Lofreq
	Blood	NA	0	F	cDNA	FLX	RC454	Lofreq
Immunised	Serum	p2	5	H	cDNA	FLX	RC454	Lofreq
	Serum	p3	5	H	cDNA	PGM	RC454	Lofreq
	Serum	p5	5	H	cDNA	PGM	RC454	Lofreq
	Serum	p15	5	H	cDNA	PGM	RC454	Lofreq
Controls	Serum	p19	3	H	cDNA	PGM	RC454	Lofreq
	Serum	p19	6	-	RNA	FLX	RC454	Lofreq
	Blood	p19	6	H	cDNA	PGM	RC454	Lofreq
	Tonsil	p19	6	F	cDNA	PGM	RC454	Lofreq
	Serum	p20	3	H	cDNA	PGM	RC454	Lofreq
	Serum	p20	6	-	RNA	FLX	RC454	Lofreq
	Blood	p20	6	H	cDNA	PGM	RC454	Lofreq
	Tonsil	p20	6	F	cDNA	PGM	RC454	Lofreq
	Serum	p21	3	H	cDNA	FLX	RC454	Lofreq
	Serum	p21	6	-	RNA	FLX	RC454	Lofreq
	Blood	p21	6	F	cDNA	FLX	RC454	Lofreq
	Tonsil	p21	6	F	cDNA	PGM	RC454	Lofreq

[1]NA: Not applicable. [2]F: Full-length RT-PCR amplicon; H: 2 x half-length RT-PCR amplicons

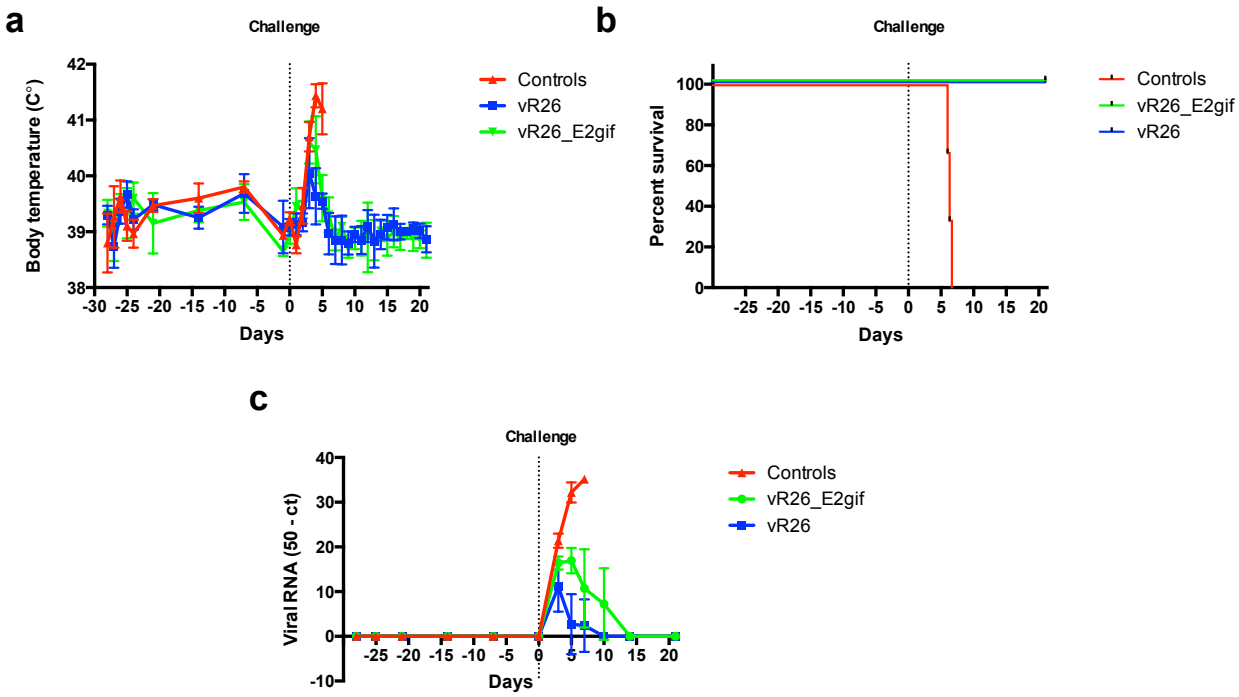


Figure 1. Summary of results from pigs immunised with vR26 and vR26_E2gif and challenge infected with highly virulent CSFV strain "Koslov" (Rasmussen et al., unpublished). a) Temperature curves. b) Pig survival curves. c) Viral RNA loads in blood measured by RT-qPCR.

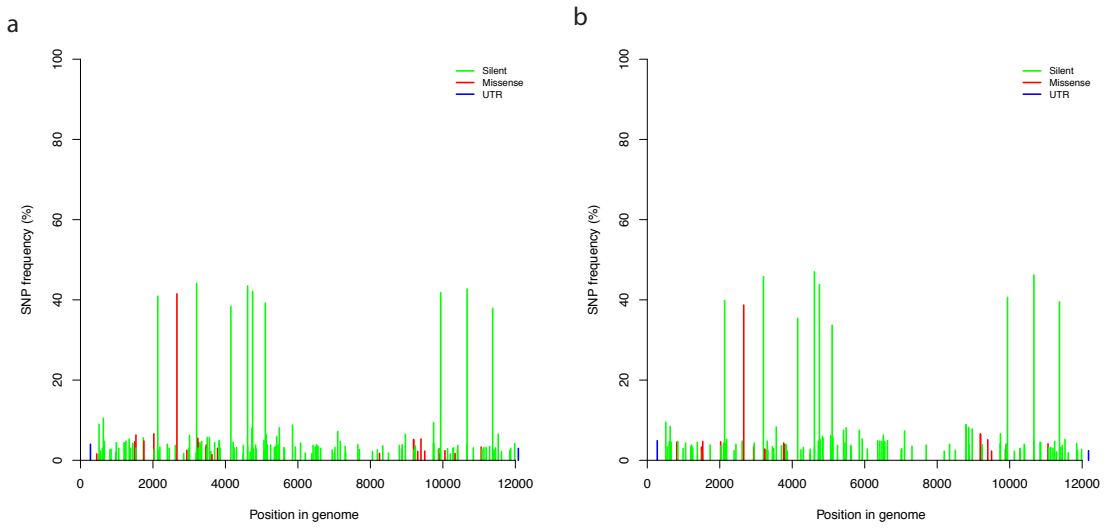
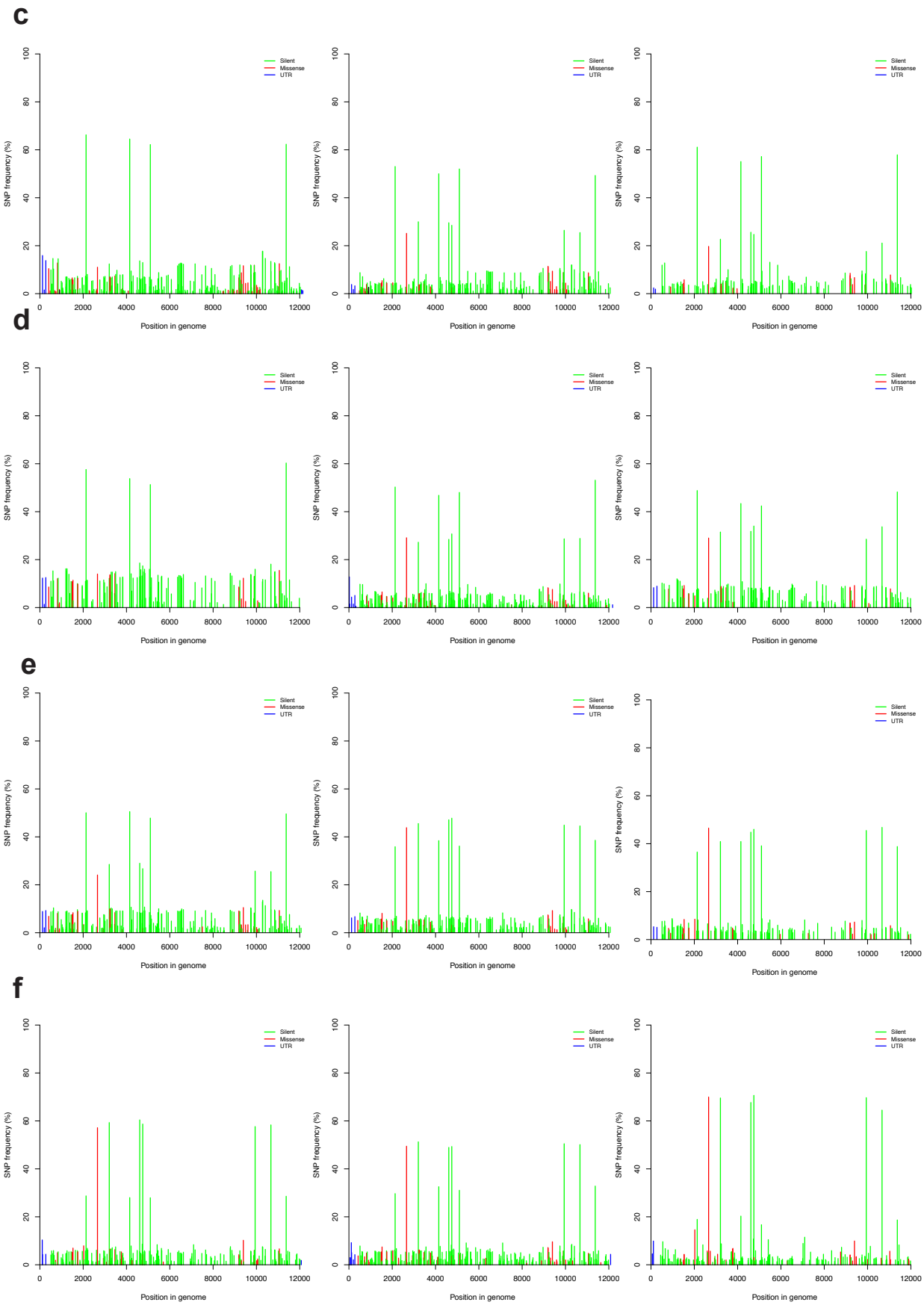


Figure 2. SNP frequency plots of all samples in this study (see supplementary table 1). a) Virus inoculum (cDNA). b) Virus inoculum (RNA). c) Serum (PID3) from control pigs (p19, p20, p21, respectively). d) Serum (PID6) from control pigs. e) Blood (PID6) from control pigs. f) Tonsils (PID6) from control pigs. g) Serum (PID5) from pig p2 immunised with vR26_E2gif. h) Serum (PID5) from pig p3 immunised with vR26_E2gif. i) Serum (PID5) from pig p5 immunised with vR26_E2gif. j) Serum (PID5) from pig p15 immunised with vR26.


c
d
e
f
Silent
Missense
UTR
SNP frequency (%)
Position in genome
0
20
40
60
80
100
0
2000
4000
6000
8000
10000
12000

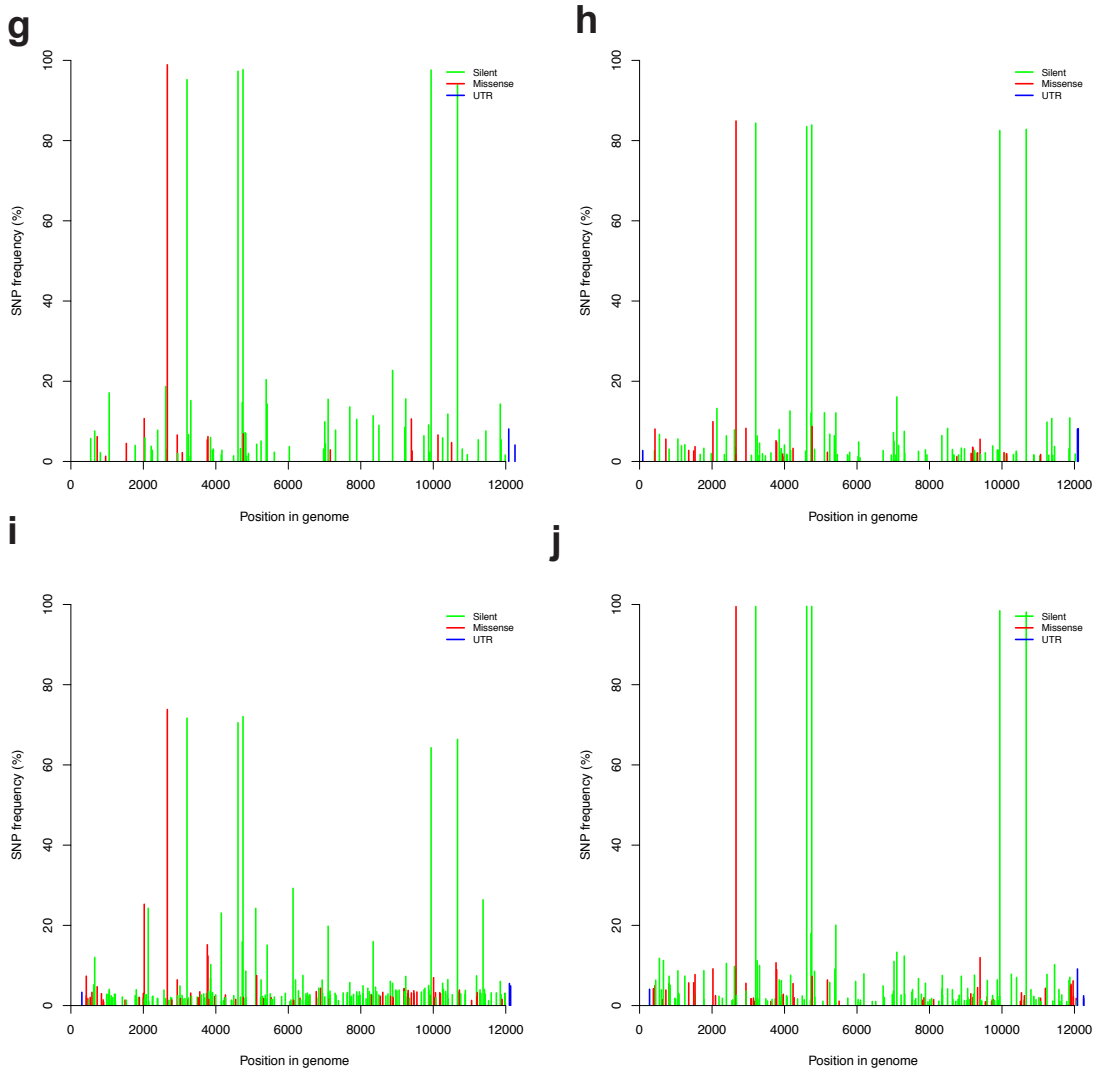


Figure 2. SNP frequency plots of all samples in this study (see supplementary table 1). a) Virus inoculum (cDNA). b) Virus inoculum (RNA). c) Serum (PID3) from control pigs (p19, p20, p21, respectively). d) Serum (PID6) from control pigs. e) Blood (PID6) from control pigs. f) Tonsils (PID6) from control pigs. g) Serum (PID5) from pig p2 immunised with vR26_E2gif. h) Serum (PID5) from pig p3 immunised with vR26_E2gif. i) Serum (PID5) from pig p5 immunised with vR26_E2gif. j) Serum (PID5) from pig p15 immunised with vR26.

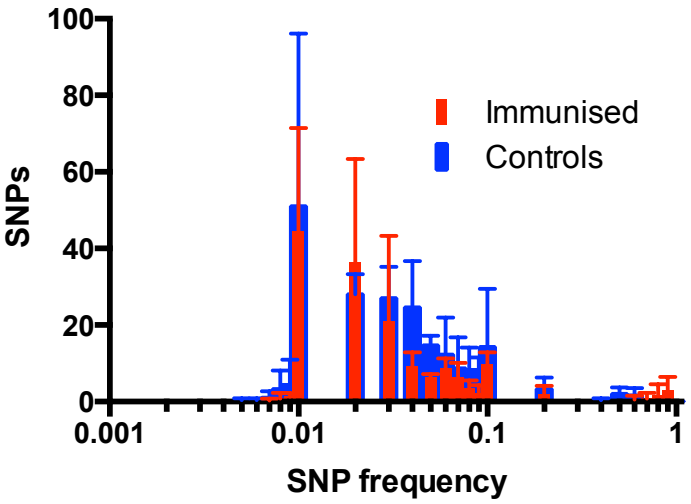


Figure 3. SNP distributions for immunised and control groups. Means ± s.d. are shown for biological replicates (n = 3 and n=4)

a

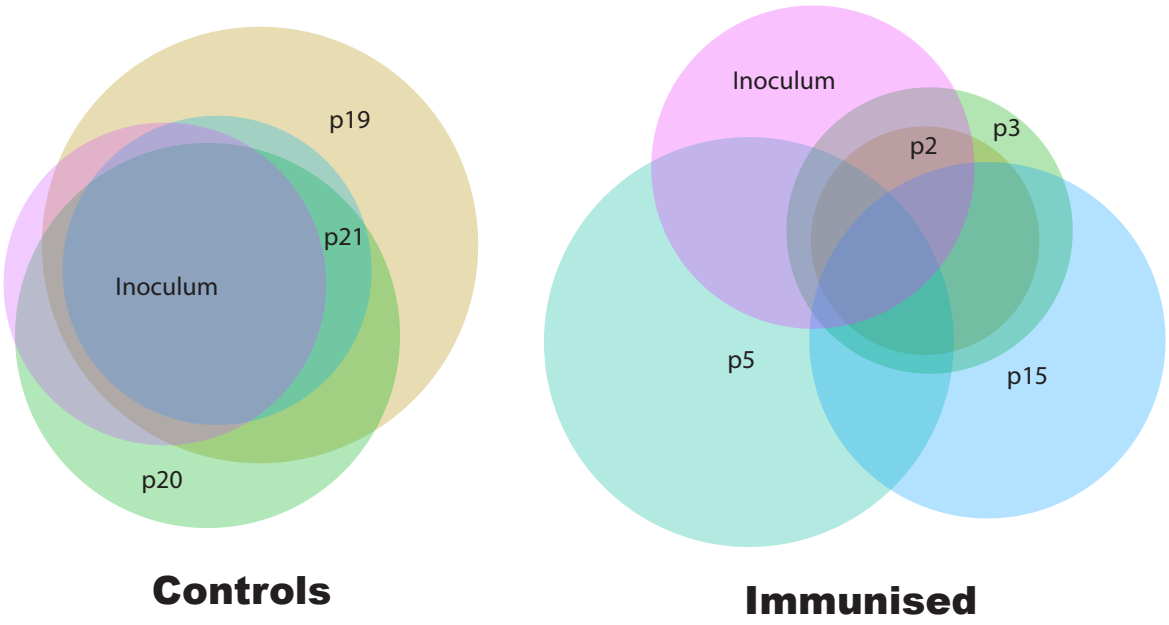


Figure 4. SNP comparison within immunised and control groups. a) Venn-diagram showing the proportion of overlapping and unique SNPs for both the controls and the immunised serum samples individually compared to the inoculum. b) Venn-diagram showing the proportion of overlapping and unique SNPs for the mock-immunised and the immunised serum samples compared to the inoculum comparing above 50% SNP intersections between animals from the same group.

b

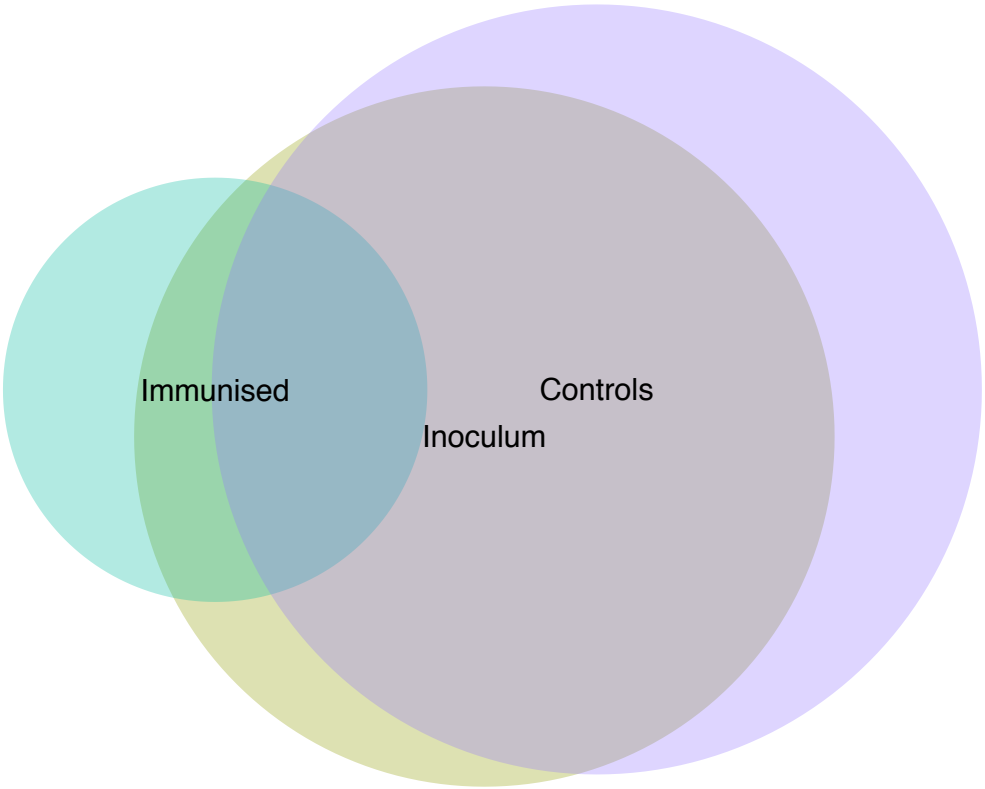


Figure 4. SNP comparison within immunised and control groups. a) Venn-diagram showing the proportion of overlapping and unique SNPs for both the controls and the immunised serum samples individually compared to the inoculum. b) Venn-diagram showing the proportion of overlapping and unique SNPs for the mock-immunised and the immunised serum samples compared to the inoculum comparing above 50% SNP intersections between animals from the same group.

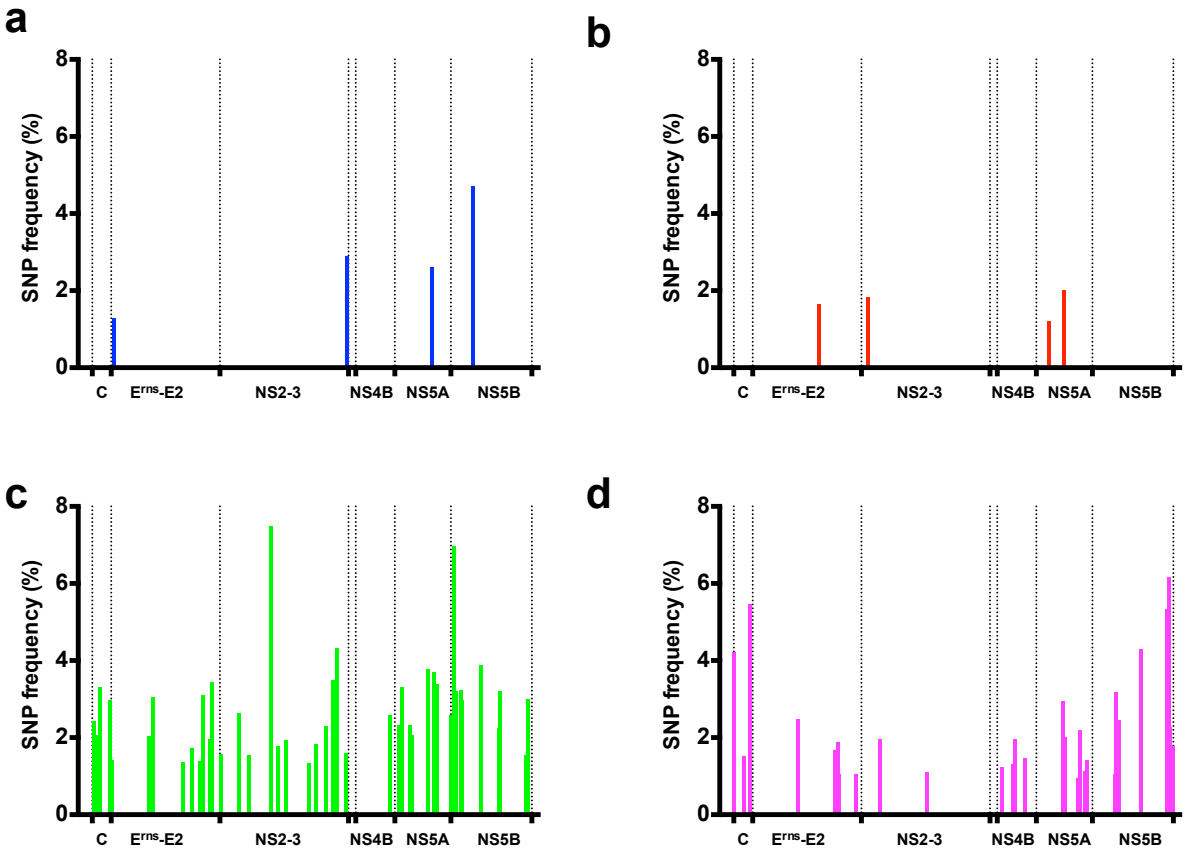


Figure 5. Unique missense SNPs detected in immunised pigs. Unique missense SNP frequency (%) distribution of each immunised pig (p2, p3, p5, p15) along the genome with protein regions depicted as vertical dashed lines. (a) p2 immunised with vR26_E2gif; (b) p3 immunised with vR26_E2gif; (c) p5 immunised with vR26_E2gif; (d) p15 immunised with vR26.

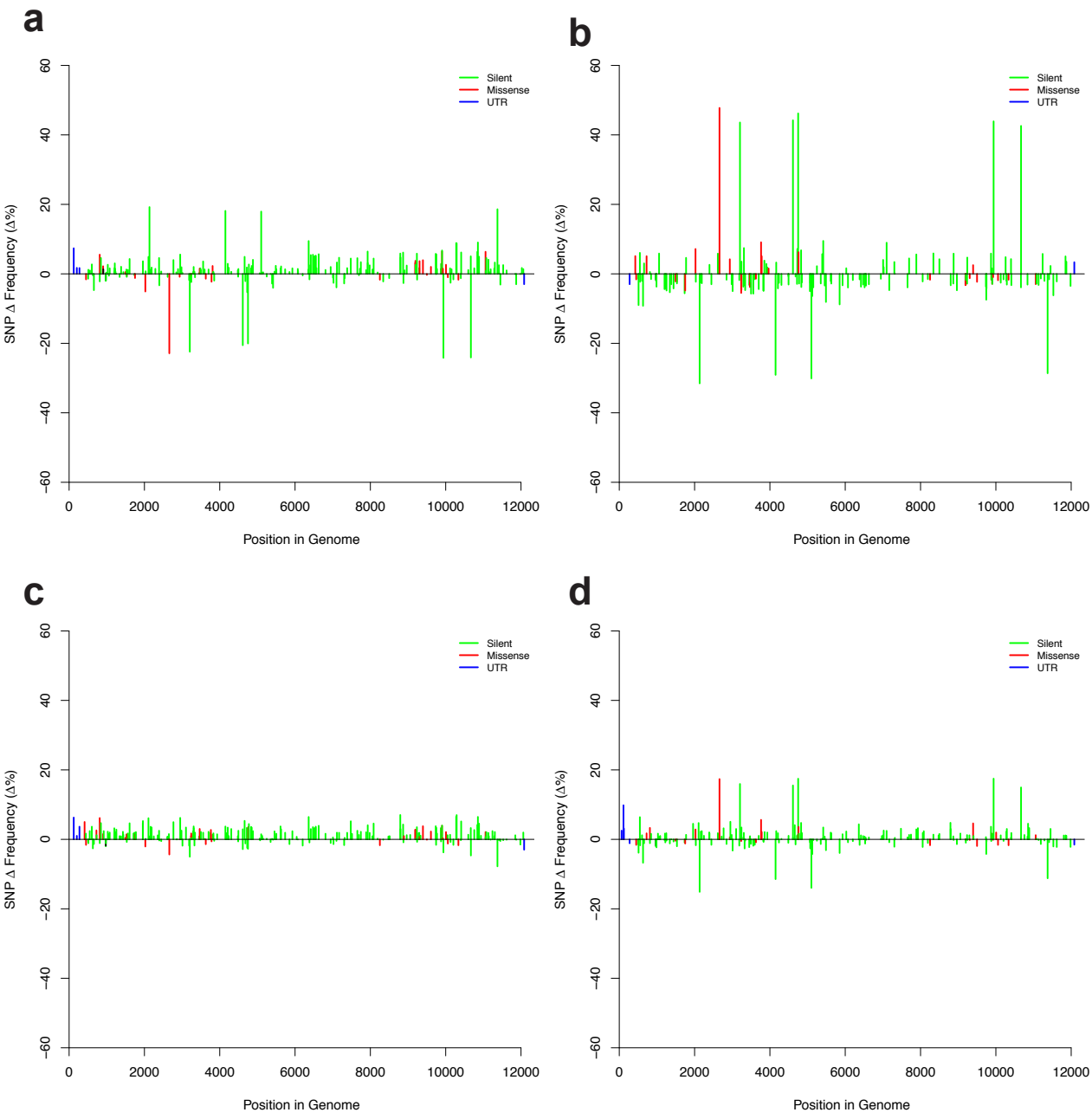


Figure 6. Relative SNP frequency distributions depicted as the mean frequency change of overlapping and unique SNPs for the immunised and control groups compared to the inoculum as Δ%. a) Serum (PID3) from control pigs. b) Serum (PID5) from immunised pigs. c) Blood (PID6) from control pigs. d) Tonsils (PID6) from control pigs.

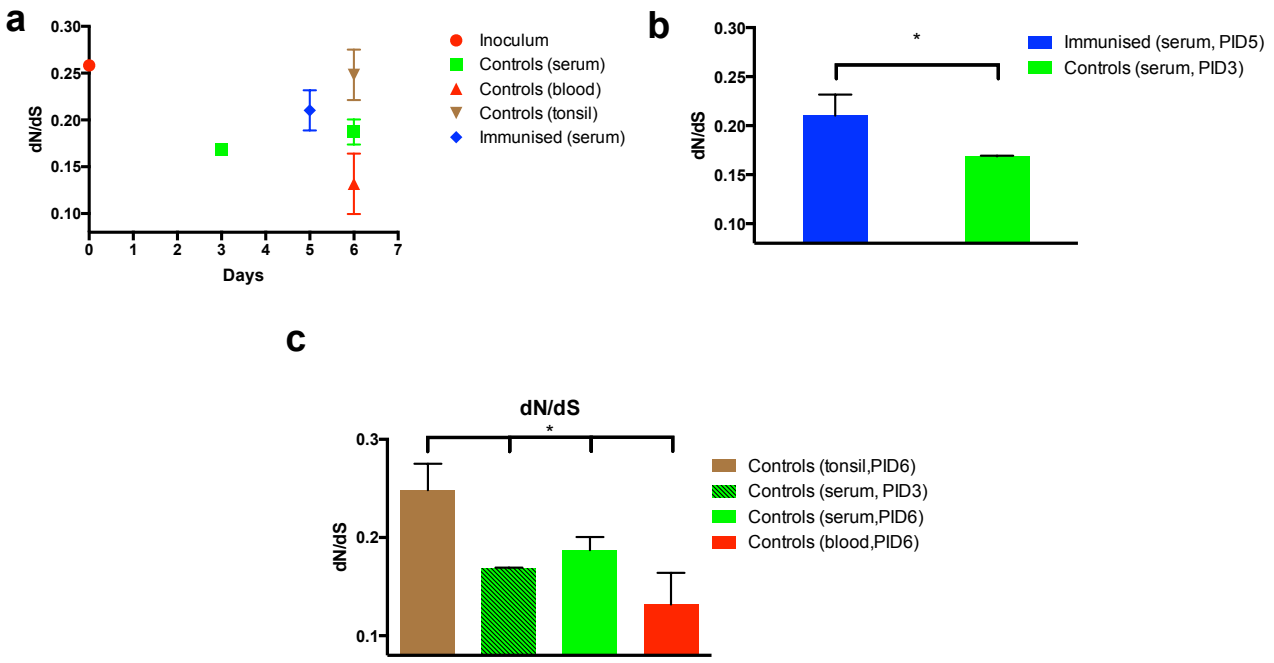


Figure 7. dN/dS analysis. a) depicts the dN/dS ratio plotted for each individual sample type during the experiment. b) The dN/dS ratio plotted for immunised and control groups. Means ± s.d. are shown for biological replicates (n = 3 and n=4). The T-test was applied to determine significance difference between the two groups (p = 0.0221). c) The dN/dS ratios plotted for the controls. Means ± s.d. are shown for biological replicates (n = 3). The T-test was applied to determine significance difference between tonsils compared to the blood and serum (PID3 and PID6) (p = 0.0087, p = 0.0069 and p = 0.0245).

Conclusions and future perspectives

Conclusions and future perspectives

RNA viruses have some of the highest mutation rates seen in nature (Drake 1993). This allows them to evade host immune responses and adapt to new hosts. Understanding this adaptive potential is crucial when making vaccines and treating infections with antiviral agents. In addition, the mutation rate allows for closely related haplotypes, with only a few differences between them, to co-exist in a RNA virus population. This viral cloud is thought to play an important role in the infectivity and virulence of the population (Lauring and Andino 2010). For CSFV many strains have been characterized with virulence ranging from avirulent to highly virulent with mortality rates of close to 100%. This raises questions about virulence that might be answered by genome sequencing and analysis of population haplotype structure. However, in the past this was only possible to study by extensive cloning and Sanger sequencing. A major breakthrough was made by the introduction of NGS sequencing into biological research. Large genomes can now be sequenced in hours at only a fraction of the cost of conventional Sanger sequencing. For virus research the area is still being developed. However, researchers are pushing the boundaries for the applications of these new technologies. For RNA viruses there are challenges that are not easily solved. The major issues are the short length reads, the error rate by both the platform and the essential reverse transcription. The short read lengths limit the ability to reconstruct haplotypes and reduce the lengths of contigs from the de novo assemblies. The error rate is not a concern for pro- and eukaryotes because it is DNA that is sequenced and no intra sample variation is expected except from heterozygous positions. The cloud of variants within an RNA virus population challenges the technologies and no definitive solution has been presented. The shear amount of data can be somewhat overwhelming and bioinformatics skills are necessary for data analysis.

This thesis has been focused on studying adapting viral populations by NGS, full-length cDNA cloning and reverse genetics. However, when we started out there was no guarantee that we would be successful in our approaches. It turned out that we were able to deep sequence from

a variety of samples using different sample preparation approaches and sequencing platforms as described above. The data analysis pipeline was set up as a collaboration between Simon Rasmussen and me facilitated by Anders Gorm Pedersen all three of us from CBS at DTU Systems Biology. I adopted and added on Simon's suggested standard tools so that it fitted RNA virus challenges. I added further improvements along the way, and went to the University of Glasgow to study with Richard Orton at Daniel Hayden's lab to immerse myself deeper into bioinformatics. The standard pipeline has been presented above and is functional for several different viruses, and can be adapted to any chosen virus. Reverse genetics and full-length cDNA cloning has also been a big part of this project which has allowed us to test hypotheses both *in vitro* and in the natural host, the pig. During my stay at the lab, we have refined a lab workflow that allows a mutagenesis idea to be implemented and tested in our reverse genetic system within two weeks. The power of this system can be seen in manuscript 3 and 4. Full-length sequencing of every single construct has proven essential and will become widely used in future studies with reverse genetics of cDNA clones. This is because the possibility of secondary mutations will potentially ruin results and must be ruled out from the start. We have sequenced all our constructs by NGS and this has the advantage of not running at least 30 Sanger sequencing reactions with both pan and specific primers for each CSFV strain (Leifer et al. 2010). The only disadvantage of NGS platform at the moment is the number of samples needed to multiplex for it to be economically feasible. Until then traditional Sanger sequencing still has relevance for its ability to quickly sequence one sample. We use Sanger sequencing for screening our cDNA constructs for correct mutagenesis before sending for full-length sequencing by NGS. Following the initial screening, we only have to send two cDNA clones of each construct for NGS and never had to sequence additional constructs to gain a 100% match to the expected sequence.

Manuscript 1 and 2 were focused on full-length sequencing of BDV strain "Gifhorn" and the CSFV strain "Bergen" respectively. De novo assemblies and subsequent mapping obtained both consensus sequences. Full-length RT-PCR was applied to both strains, which also gave deep sequencing information that explained the discrepancies between the full-length Bergen consensus and the short sequences published earlier for SNPs in the population. This was also the first genotype 2.2 strain to be fully sequenced. The publication of these two manuscripts proved that our approach could be used for any pestivirus genome and access to more

pestivirus strains could rapidly increase the number of full genome pestivirus sequences in the databases.

One of the most virulent CSFV strains described is the “Koslov” isolate with mortality rates of 100% in both domestic pigs and in wild boars (Bartak and Greiser-Wilke 2000; Blome et al. 2012). Previously, no functional cDNA clone had been reported. A cDNA clone should prove an important tool in understanding of the virulence for CSFV. In the lab, we had produced four full-length clones of which none were infectious. Instead of abandoning the experiment, we full-length sequenced all four and analyzed the sequences. Two of the sequences had indels in their reading frames and the other two had several missense mutations. We suggested that bringing the cDNA closer to the consensus sequence at the coding level would increase the possibility of the clone to be infectious. The first site-directed mutagenesis step left us with a cDNA with only four missense mutations compared to the consensus (Kos_4aa). This construct turned out to be infectious and a further three rounds of mutagenesis gave rise to rescued viruses; vKos_3aa, vKos_2aa and vKos all of which grew like the virus rescued from the parental (uncloned) cDNA. Subsequently, vKos_3aa and vKos were tested in infection experiments in pigs and showed clear differences in virulence with the vKos being as highly virulent as the parental Koslov virus and the vKos_3aa being significantly less virulent. Deep sequencing of serum samples revealed no adaptations in the inoculated pigs in the vKos group, which was not the case for vKos_3aa where all inoculated pigs reverted to the parental state at position C996T in the core protein. These molecular micro-evolution studies gave us hints for future experiments into the significance of the S763L missense mutation. A modified version of the Kos was produced, termed Kos_S763L, and tested in pigs and showed similar virulence (data not published). This work emphasized the importance of these virus adaptation studies and that virulence cannot at this stage be determined in vitro, but animal infection experiments are still needed to determine virulence.

During the analysis described in manuscript 3 it was evident that some of the mutations in the cDNA clones could be found as low frequency SNPs in deep sequencing of the parental population. This gave rise to the idea that a lot of full-length cDNA clones from the same population must represent the major haplotypes. However, we only had four cDNA clones in

manuscript 3, which was not enough. Manuscript 4 is an in depth study of the CSFV "Roesrath" isolate population. Here we managed to get 84 cDNA clones in one go. Only 12% were infectious and the majority of RNAs transcribed from the cDNA were non-functional. This should be due to the error prone nature of the RdRp incorporating mutations randomly over the genome. A significant difference in missense mutations was observed between the infectious and the non-functional cDNAs. Thus if a genome had accumulated many missense mutations it is most likely non-functional. NGS allowed deep sequencing of the uncloned RT-PCR product and most major groups of the cDNA phylogenies were confirmed by the SNP analysis. In addition, having the phylogeny of the cDNAs allowed us to benchmark the NGS haplotype predictors, which revealed that they were only producing reliable predictions from viruses of extremely heterogeneous populations (Di Giallonardo et al. 2013; Giallonardo et al. 2014) and not from our relative homogeneous population. A major limiting factor is the short read length. The PacBio platform (Pacific Biosciences) has the potential of giving reads of up to 10 kb that could complement normal NGS and bridge the long gaps between SNPs and allow better haplotype reconstruction. However, the machine requires several µg of high quality template DNA to work. So it might be possible with RT-PCR amplification but not by second strand cDNA synthesis to produce enough material for PacBio sequencing.

The ancestral reconstruction of internal nodes proved successful in producing fully functional viruses. When tested, differences were observed between the vRos and the vRos_S1359N_A2668T *in vitro* and *in vivo*. vRos_S1359N_A2668T replicated faster *in vitro* and the vRos had a virulence comparable to the Roesrath isolate, while the vRos_S1359N_A2668T had an attenuated phenotype *in vivo*. Deep sequencing of the original isolate (passage 2) revealed that the S1359N_A2668T SNPs were not detectable in the population. Taken together these SNPs are most likely cell culture adaptations. Further cell culture studies can reveal if these mutations work in concert or if it is only one that is necessary for increased replication speed in cell culture.

Manuscript 5 was a study of the CSFV adaptation potential to the highly selective pressure imposed by vaccination. This study included NGS data from immunised and control pigs challenge infected by the CSFV strain "Koslov". Several sequencing sample preparation

approaches were successful and we were able to amplify and deep sequence from low viral load samples from the immunised pigs. This data allowed us to take a different approach than with normal vaccine trials by studying the "Koslov" population under selection pressure and to look for adaptation and possible escape mutants. In the analysis we learned that two major haplotypes seem to dominate the population, One with 4 silent SNPs and the other with 5 silent SNPs and one missense SNP (S763L). The consensus sequence of the inoculum can be seen as a distribution rather than actually being represented on the majority of the viral genomes. We found a change in population structure for all immunized animals compared to the controls. Especially, the S763L mutation also discussed in manuscript 3 seemed to be almost fixed at day 5 in the immunised pigs. The controls had a decrease in the serum of that particular SNP. However, in the tonsil samples from the control group an increase of S763L was observed pointing towards this haplotype replicating faster in the tonsils. However, the S763 haplotype must then replicate faster elsewhere in the pig to maintain both haplotypes. So after vaccination the "Koslov" virus mainly replicates in the tonsils thereby maintaining the S763L haplotype and that is why it was found almost fixed in serum. In contrast, when no vaccination has been performed the population is allowed to replicate without selection pressure of the immunisation leading to the maintenance of the S763 haplotype. The data clearly hints at tropism effect of this SNP, which has been shown to be part of a potential epitope on the E2 surface protein (Chang et al. 2012). Further animal studies have been planned and performed within our group into this epitope with E2 modified versions of the Kos cDNA from manuscript 3. From those *in vivo* experiments different tissue samples (brain, lymph, spleen, etc) were obtained to look for different virus haplotype structures as seen in deep sequencing of serum and tonsils and observed in other viruses (Wright et al. 2011). The results from these experiments are still being processed but initial analysis shows that this epitope is important for virulence and modification can increase virulence (data not shown). We also observed several low frequency missense SNPs (1-10%) in the immunised pigs that could point to adaptation or possible escape mutants. Further analysis might reveal if they are positively selected. The dN/dS analysis did reveal that the immunised animals to be under slightly more positive selection in accordance the strong selection imposed by the vaccine. However, the analysis is still being refined and will hopefully in the future allow looking at individual sites for positive selection. Some of the challenges lie in the overestimation, based on the lack of knowledge, of the populations underlying the phylogeny obtained from the NGS data. If a mutation has occurred deep in the phylogeny once and is positively selected or

hitchhiked together on a haplotype with another positively selected mutation then the calculations will overestimate the dN/dS ratio. This can be solved, in principle, by *in silico* haplotype reconstruction, but as shown above the performance of these tools are not reliable enough at the moment. Another option we are exploring is to reduce the phylogeny effect by a weight matrix thereby making the phylogeny more star-like. This is still ongoing and will hopefully increase the credibility of these dN/dS ratio calculations together with a Bayesian statistical approach to the true likelihood in which we can trust.

Other work

During the three years of my PhD study I have been involved in several projects that included national and international collaboration. Some of this work has not ended up in the thesis and is worth mentioning. In 2012, I co-supervised Abdou Mohammed Nagy an Egyptian student during his project at Lindholm. The goal was to sequence Danish BVDV isolates to look for genetic heritage and adaptation. The lab work was designed and supervised by me with Abdou at the bench; I also undertook the subsequent sequence analysis. This work was finished up in a publication (Nagy et al. 2013). Another student Nana Baluhla from University of Copenhagen came and performed a project in our group in the early winter months of 2013. She was involved in the generation of the BDV "Gifhorn" cDNA clone. The Gifhorn cDNA clone has since been modified to completely resemble the sequence published in manuscript 1 except for one silent mutation. The cDNA clone is fully functional and replicates as the parental isolate in ovine SFT-R cells. We have done further studies into the BDV "Gifhorn" strain by full-length sequencing of a sheep isolate that differs significantly from the published pig isolate. A cell culture host adaptation study has been performed in primary cells from both lambs and pigs. Results are still being analysed. A big part of this project has been in NGS data analysis. During this time I have been involved with data analysis in Lise Kviisgaards PhD project at DTU Vet in Copenhagen. She was involved in full-length sequencing of porcine reproductive respiratory syndrome virus (PRRSV) from both the European and American genotype. This involved the FLX, Ion PGM and the Illumina data and the comparison of platforms with me as co-author (Kvisgaard et al. 2013). Emma Hagberg, a PhD student at DTU Systems Biology, has performed NGS full-length sequencing of Aleutian disease virus (ADV). I

have helped her design sample preparation and she has adopted part of my NGS data pipeline that is producing good results. In addition, I have been involved in sequencing during PhD Peter Christian Risagers project that focused on virus replication using reporter replicons (Risager et al. 2013). Another aspect of his project was the chimeric vaccine C-strain replicon with the highly virulent "Koslov" RdRp NS5B inserted, which speeded up replication. These promising results were later backed by the work of master student Jonas Kjær during his thesis that was co-supervised by me. We set up a replication assay also used in manuscript 4 and I devised the mutagenesis strategies. Finally, I have been involved in NGS analysis of a virulence study of vPader10 and IRES mutants derived from the CSFV strain "Paderborn" cDNA clone (Rasmussen et al. 2010; Friis et al. 2012). This deep sequencing of inoculums and blood samples from the animals with severe symptoms of CSF have revealed mutations that were reversions to the wild type "Paderborn" strain.

References

Literature Cited

Acevedo A, Brodsky L, Andino R. 2014. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature 505:686-690.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bartak P, Greiser-Wilke I. 2000. Genetic typing of classical swine fever virus isolates from the territory of the czech republic. Vet. Microbiol. 77:59-70.

Beer M, Reimann I, Hoffmann B, Depner K. 2007. Novel marker vaccines against classical swine fever. Vaccine 25:5665-5670.

Blome S, Aebischer A, Lange E, Hofmann M, Leifer I, Loeffen W, Koenen F, Beer M. 2012. Comparative evaluation of live marker vaccine candidates "CP7_E2alf" and "flc11" along with C-strain "riems" after oral vaccination. Vet. Microbiol. 158:42-59.

Cai W, Pei J, Grishin NV. 2004. Reconstruction of ancestral protein sequences and its applications. BMC Evol. Biol. 4:33.

Chang CY, Huang CC, Deng MC, Huang YL, Lin YJ, Liu HM, Lin YL, Wang FI. 2012. Antigenic mimicking with cysteine-based cyclized peptides reveals a previously unknown antigenic determinant on E2 glycoprotein of classical swine fever virus. Virus Res. 163:190-196.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff:

SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80-92.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. (x co-authors. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158.

Di Giallonardo F, Zagordi O, Duport Y, Leemann C, Joos B, Kunzli-Gontarczyk M, Bruggmann R, Beerenwinkel N, Gunthard HF, Metzner KJ. 2013. Next-generation sequencing of HIV-1 RNA genomes: Determination of error rates and minimizing artificial recombination. PLoS One 8:e74249.

Drake JW. 1993. Rates of spontaneous mutation among RNA viruses. Proc. Natl. Acad. Sci. U. S. A. 90:4171-4175.

El Omari K, Iourin O, Harlos K, Grimes JM, Stuart DI. 2013. Structure of a pestivirus envelope glycoprotein E2 clarifies its role in cell entry. Cell. Rep. 3:30-35.

Fahnøe U, Lohse L, Becher P, Rasmussen TB. 2014a. Complete genome sequence of classical swine fever virus genotype 2.2 strain bergen. Genome Announc 2:10.1128/genomeA.00483-14.

Fahnøe U, Höper D, Schirrmeier H, Beer M, Rasmussen TB. 2014b. Complete genome sequence of border disease virus genotype 3 strain gifhorn. Genome Announc 2:10.1128/genomeA.01142-13.

Fahnøe U, Pedersen AG, Risager PC, Nielsen J, Belsham GJ, Höper D, Beer M, Rasmussen TB. 2014c. Rescue of the highly virulent classical swine fever virus strain "Koslov" from cloned

cDNA and first insights into genome variations relevant for virulence. Virology 468-470C:379-387.

Floegel-Niesmann G, Blome S, Gerss-Dulmer H, Bunzenthal C, Moennig V. 2009. Virulence of classical swine fever virus isolates from Europe and other areas during 1996 until 2007. Vet. Microbiol. 139:165-169.

Friis MB, Rasmussen TB, Belsham GJ. 2012. Modulation of translation initiation efficiency in classical swine fever virus. J. Virol. 86:8681-8692.

Giallonardo FD, Töpfer A, Rey M, Prabhakaran S, Duport Y, Leemann C, Schmutz S, Campbell NK, Joos B, Lecca MR et al. 2014. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res. 42:e115.

Gladue DP, O'Donnell V, Fernandez-Sainz IJ, Fletcher P, Baker-Branstetter R, Holinka LG, Sanford B, Carlson J, Lu Z, Borca MV. 2014. Interaction of structural core protein of classical swine fever virus with endoplasmic reticulum-associated degradation pathway protein OS9. Virology 460-461:173-179.

Gottipati K, Acholi S, Ruggli N, Choi KH. 2014. Autocatalytic activity and substrate specificity of the pestivirus N-terminal protease Npro. Virology 452-453:303-309.

Gottipati K, Ruggli N, Gerber M, Tratschin JD, Benning M, Bellamy H, Choi KH. 2013. The structure of classical swine fever virus N(pro): A novel cysteine autoprotease and zinc-binding protein involved in subversion of type I interferon induction. PLoS Pathog. 9:e1003704.

Gullberg M, Tolf C, Jonsson N, Mulders MN, Savolainen-Kopra C, Hovi T, Van Ranst M, Lemey P, Hafenstein S, Lindberg AM. 2010. Characterization of a putative ancestor of coxsackievirus B5. J. Virol. 84:9695-9708.

Hashem Y, des Georges A, Dhote V, Langlois R, Liao HY, Grassucci RA, Pestova TV, Hellen CU, Frank J. 2013. Hepatitis-C-virus-like internal ribosome entry sites displace eIF3 to gain access to the 40S subunit. Nature 503:539-543.

Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, Berlin AM, Malboeuf CM, Ryan EM, Gnerre S et al. 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 8:e1002529.

Hoffmann B, Scheuch M, Höper D, Jungblut R, Holsteg M, Schirrmeier H, Eschbaumer M, Goller KV, Wernike K, Fischer M et al. (x co-authors. 2012. Novel orthobunyavirus in cattle, europe, 2011. Emerg. Infect. Dis. 18:469-472.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22:568-576.

Kvisgaard LK, Hjulsager CK, Fahnøe U, Breum SO, Ait-Ali T, Larsen LE. 2013. A fast and robust method for full genome sequencing of porcine reproductive and respiratory syndrome virus (PRRSV) type 1 and type 2. J. Virol. Methods 193:697-705.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. Nat. Methods 9:357-359.

Lauring AS, Andino R. 2010. Quasispecies theory and the behavior of RNA viruses. PLoS Pathog. 6:e1001005.

Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. 2014. MOSAIK: A hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One 9:e90581.

Leifer I, Hoffmann B, Höper D, Bruun Rasmussen T, Blome S, Strebelow G, Horeth-Bontgen D, Staubach C, Beer M. 2010. Molecular epidemiology of current classical swine fever virus isolates of wild boar in germany. J. Gen. Virol. 91:2687-2697.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. arXiv:1303.3997v2. http://arxiv.org/abs/1303.3997v2.

Li Y, Wang J, Kanai R, Modis Y. 2013. Crystal structure of glycoprotein E2 from bovine viral diarrhea virus. Proc. Natl. Acad. Sci. U. S. A. 110:6805-6810.

Logan G, Freimanis GL, King DJ, Valdazo-Gonzalez B, Bachanek-Bankowska K, Sanderson ND, Knowles NJ, King DP, Cottam EM. 2014. A universal protocol to generate consensus level genome sequences for foot-and-mouth disease virus and other positive-sense polyadenylated RNA viruses using the illumina MiSeq. BMC Genomics 15:828.

Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads EMBnetjournal, North America 2011, 17 http://journal.embnet.org/index.php/embnetjournal/article/view/200/479 webcite

Maurer K, Krey T, Moennig V, Thiel HJ, Rumenapf T. 2004. CD46 is a cellular receptor for bovine viral diarrhea virus. J. Virol. 78:1792-1799.

Mayer D, Thayer TM, Hofmann MA, Tratschin JD. 2003. Establishment and characterisation of two cDNA-derived strains of classical swine fever virus, one highly virulent and one avirulent. Virus Res. 98:105-116.

Meyers G, Thiel HJ, Rumenapf T. 1996. Classical swine fever virus: Recovery of infectious viruses from cDNA constructs and generation of recombinant cytopathogenic defective interfering particles. J. Virol. 70:1588-1595.

Mittelholzer C, Moser C, Tratschin JD, Hofmann MA. 1997. Generation of cytopathogenic subgenomic RNA of classical swine fever virus in persistently infected porcine cell lines. Virus Res. 51:125-137.

Mittelholzer1 C, Moser2 C, Tratschin JD, Hofmann MA. 2000. Analysis of classical swine fever virus replication kinetics allows differentiation of highly virulent from avirulent strains. Vet. Microbiol. 74:293-308.

Moormann RJ, van Gennip HG, Miedema GK, Hulst MM, van Rijn PA. 1996. Infectious RNA transcribed from an engineered full-length cDNA template of the genome of a pestivirus. J. Virol. 70:763-770.

Morelli MJ, Wright CF, Knowles NJ, Juleff N, Paton DJ, King DP, Haydon DT. 2013. Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. Vet. Res. 44:12-9716-44-12.

Moser C, Stettler P, Tratschin JD, Hofmann MA. 1999. Cytopathogenic and noncytopathogenic RNA replicons of classical swine fever virus. J. Virol. 73:7787-7794.

Murray CL, Jones CT, Rice CM. 2008. Architects of assembly: Roles of flaviviridae non-structural proteins in virion morphogenesis. Nat. Rev. Microbiol. 6:699-708.

Nagy A, Fahnøe U, Rasmussen TB, Uttenthal A. 2013. Studies on genetic diversity of bovine viral diarrhea viruses in danish cattle herds. Virus Genes 48: 376-380

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Postel A, Moennig V, Becher P. 2013. Classical swine fever in europe--the current situation. Berl. Munch. Tierarztl. Wochenschr. 126:468-475.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842.

Rasmussen TB, Reimann I, Uttenthal A, Leifer I, Depner K, Schirrmeier H, Beer M. 2010. Generation of recombinant pestiviruses using a full-genome amplification strategy. Vet. Microbiol. 142:13-17.

Rasmussen TB, Risager PC, Fahnøe U, Friis MB, Belsham GJ, Höper D, Reimann I, Beer M. 2013. Efficient generation of recombinant RNA viruses using targeted recombination-mediated mutagenesis of bacterial artificial chromosomes containing full-length cDNA. BMC Genomics 14:819.

Risager PC, Fahnøe U, Gullberg M, Rasmussen TB, Belsham GJ. 2013. Analysis of classical swine fever virus RNA replication determinants using replicons. J. Gen. Virol. 94:1739-1748.

Risatti GR, Holinka LG, Carrillo C, Kutish GF, Lu Z, Tulman ER, Sainz IF, Borca MV. 2006. Identification of a novel virulence determinant within the E2 structural glycoprotein of classical swine fever virus. Virology 355:94-101.

Ronquist F. 2004. Bayesian inference of character evolution. Trends Ecol. Evol. 19:475-481.

Ruggli N, Tratschin JD, Mittelholzer C, Hofmann MA. 1996. Nucleotide sequence of classical swine fever virus strain alfort/187 and transcription of infectious RNA from stably cloned full-length cDNA. J. Virol. 70:3478-3487.

Salmela L, Schroder J. 2011. Correcting errors in short reads by multiple alignments. Bioinformatics 27:1455-1461.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863-864.

Tamura T, Sakoda Y, Yoshino F, Nomura T, Yamamoto N, Sato Y, Okamatsu M, Ruggli N, Kida H. 2012. Selection of classical swine fever virus with enhanced pathogenicity reveals synergistic virulence determinants in E2 and NS4B. J. Virol. 86:8602-8613.

Töpfer A, Höper D, Blome S, Beer M, Beerenwinkel N, Ruggli N, Leifer I. 2013. Sequencing approach to analyze the role of quasispecies for classical swine fever. Virology 438:14-19.

Van Gennip HG, Vlot AC, Hulst MM, De Smit AJ, Moormann RJ. 2004. Determinants of virulence of classical swine fever virus strain brescia. J. Virol. 78:8812-8823.

Williams PD, Pollock DD, Blackburne BP, Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput. Biol. 2:e69.

Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40:11189-11201.

Wright CF, Morelli MJ, Thebaud G, Knowles NJ, Herzyk P, Paton DJ, Haydon DT, King DP. 2011. Beyond the consensus: Dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. J. Virol. 85:2266-2275.

Yang X, Charlebois P, Macalalad A, Henn MR, Zody MC. 2013. V-phaser 2: Variant inference for viral populations. BMC Genomics 14:674-2164-14-674.

Yang Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555-556.

Zeng J, Wang H, Xie X, Li C, Zhou G, Yang D, Yu L. 2014. Ribavirin-resistant variants of foot-and-mouth disease virus: The effect of restricted quasispecies diversity on viral virulence. J. Virol. 88:4008-4020.

National Veterinary Institute
Technical University of Denmark

Bülowsvej 27
1870 Frederiksberg C

www.vet.dtu.dk